UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Prostate Cancer Biochemical Recurrence Prediction After Radical Prostatectomy Using Machine Learning Analysis of Histopathology

Carolina Alexandra Carrapiço Seabra

Mestrado Integrado em Engenharia Biomédica e Biofísica
Perfil em Radiações em Diagnóstico e Terapia

Dissertação orientada por:
Dra. Raquel Conceição
Dr. Nickolas Papanikolaou

2019

Acknowledgements

First of all, I would to thank Doctor Nickolas Papanikolaou, for taking me in the Computational Clinical Imaging group at the Champalimaud Foundation. I am thankful to be part of such an important research institution and experience a truly clinical research environment.

My deepest appreciation is extended to Professor Raquel Cruz Conceição. I am grateful for all her support, regular and prompt feedback, motivation, guidance and optimism throughout all the project. I am thankful for her aspiring guidance, invaluably constructive criticism and friendly advice. Apart from her supervisor role, Professor Raquel was also a friend who kindly dealt with my "struggle" and cheered me up to continue this work. She is an inspiration for me, not only because of her ability to show her passion for the work she does, her immense knowledge and dedication to everything she is involved in, but mainly because of her kind, caring and assertive personality. It is my deepest belief, I would not have reach far without her.

Special mention goes to Prof Dr Doctor António Beltran, without whom I would still be trying to learn how to differentiate between positive and negative images. Although we spent a lot of the time doing a laborious task as checking thousands of images to perform the annotations, Prof. Dr. António was always in a good mood, especially when "PIN's guapos" came into view. It was such an opportunity for me to work in this project with him. The longer I spent with him, the more impressed I felt with his expertise, excitement and professionalism.

I would also like to express my gratitude to the rest of the group, especially João Santinha. He was always willing to generously share his time, providing me invaluable help with computer/programming issues. Our discussions over my problems and findings allowed me to choose the right direction and successfully complete my dissertation. A big thanks is also extended to José Moreia for all his help and support regarding deep learning and NiftyNet issues. Without his immense expertise I would have not been able to successfully develop my models. A special mention goes to author of the *"Parallel Genetic Algorithms For Financial Pattern Discovery Using Gpus"* bestseller. João Baúto saved me from loosing all my data a couple of times and played a decisive role, as far as the training of my models are concerned. The long hours he was willing to spent with me waiting for results were everything but boring. His intelligence, willingness to help others, sense of humor, kindness make him one of the most amazing persons I had the pleasure to meet, during the last year.

This journey would not have been the same without my dearest friend Mónica. Thank you for support, laughter, patience, silliness, encouragement, conversations, hugs, cookies... Sara for your happy mood and all the end of day skate rides. Neusa, Laeticia and Honório, I am very thankful to have you as my friends and this journey would not be the same without your friendliness. Last but not least important, I owe more than thanks to my family, for their support and encouragement throughout my life. My mother who gives me unconditional support and guides me through the difficult situations. *All that I am or hope to be I owe to you.*

Resumo

O cancro de próstata é o segundo tipo de cancro com maior prevalência nos homens, em todo o mundo. A deteção inicial desta doença ocorre, geralmente, durante exames e consultas de rotina, quando níveis aumentados do antigénio prostático específico e/ou um exame retal anormal são descobertos. Contudo, apenas a avaliação histopatológica, inicialmente baseada em amostras extraídas através de uma biópsia, é capaz de fornecer um diagnóstico definitivo, permitindo não só orientar o tratamento do doente, assim como o processo de tomada de decisão associado. Após esta avaliação inicial, se o doente for diagnosticado com cancro da próstata localizado, ou seja, um tumor confinado à próstata, o tratamento mais adotado é a prostatectomia radical. Este último é o tratamento padrão utilizado quando se pretende uma terapia com curativa. A vantagem da técnica de prostatectomia radical é que, toda a próstata, onde o tumor se encontra confinado, é removida cirurgicamente, auxiliando na redução do risco de metástases.

A posterior análise da peça prostática permite ao patologista avaliar diversas características tumorais, determinantes no prognóstico do doente. Para este fim, o microscópio tem sido a principal ferramenta utilizada, uma vez que proporciona imagens ao vivo com uma ótima resolução. No entanto, desde a introdução do primeiro sistema automatizado de digitalização de lâminas em imagens de alta resolução, o interesse da comunidade de anatomia patológica em explorar este tipo de métodos para diferentes aplicações tem crescido exponencialmente. O potencial desta nova área não é, todavia, a simples transferência de uma imagem da lâmina de vidro para um monitor, nem tão pouco a flexibilidade de distribuição e modificação da própria imagem digital, mas sim, a possibilidade de aprimorar a avaliação do patologista com informações e inteligência que não podem ser detetados pela análise humana.

Por conseguinte, a implementação de algoritmos de aprendizagem automática capazes de executar tarefas como deteção, classificação e segmentação de imagens digitais histopatológicas é, finalmente, possível. Estes métodos de análise automatizada permitem explorar todo o panorama morfológico do tumor e dos seus elementos mais invasivos presentes, capturando, por exemplo, a orientação nuclear, a textura, a forma e a arquitetura. A complexidade e densidade inerente a este tipo de imagens, oferecem uma abundância de informação, ideal para estimular e promover o desenvolvimento de algoritmos baseados em *deep learning*. No que diz respeito ao cancro da próstata, os algoritmos desenvolvidos visam apoiar as avaliações efetuados pelos patologistas, nomeadamente, estadiamento e classificação, sendo, portanto, focados no sistema de classificação utilizada para a próstata - o *Gleason score*. Contudo, problemáticas alternativas podem também beneficiar da aplicação destas técnicas, nomeadamente métodos capazes de distinguir imagens que apenas contêm tecido benigno de imagens em que tumor esteja presente. Por outro lado, modelos capazes de prever a recidiva bioquímica de cancro da próstata permitiriam aos médicos modificar estratégias de tratamento e pós-tratamento, a fim de equilibrar benefícios e efeitos adversos de um terapia específica. A previsão de recidiva permite, também, que os pacientes

escolham com responsabilidade os diferentes tratamentos e estratégias que lhes são propostos pelos médicos, possibilitando, em última análise, uma maior sua satisfação após o tratamento.

Desta forma, com o objetivo de explorar os referidos problemas, a presente dissertação apresenta procedimentos de recolha, processamento e anotação de dados, que permitiram a criação de uma base digital de dados histológicos anotados da próstata. Com base nestes dados dois modelos distintos de *deep learning*, especificamente *Convolutional Neural Networks* foram desenvolvidos. O modelo I propõe a identificação de cancro da próstata e diferenciação entre tumor e tecido benigno. O modelo II pretende prever a condição de recidiva bioquímica do cancro de próstata, para um período de tempo posterior à cirurgia em dois anos.

Relativamente ao desenvolvimento da base de dados, 200 casos de cancro da próstata, tratados através de prostatectomia radical, foram selecionados. As lâminas correspondentes à lesão índice, ou seja, à lesão principal, foram identificadas e, apenas estas foram incluídas na amostra final. A adoção desta abordagem deveu-se ao facto de que cada peça origina entre 15-45 lâminas, sendo que a maioria não contém tumor. Por outro lado, dado o período de tempo para a realização de todo este projeto, seria inviável a utilização de todas as lâminas. Assim, as lâminas selecionadas foram digitalizadas e processadas. Uma técnica de normalização de contraste foi aplicada, de forma a uniformizar as cores das diferentes imagens digitais, evitando uma elevada variabilidade de contraste e cor que advém da utilização de diferentes protocolos de cor, bem como da própria digitalização da lâmina. As imagens histológicas digitalizadas e normalizadas foram posteriormente divididas em imagens mais pequenas, isto é, subimagens, uma vez que desta forma existe uma otimização da extração de características por parte dos algoritmos. Estas subimagens foram individualmente visualizadas e anotadas, originando um total de cerca de 160,000 subimagens, correspondentes aos 200 casos diferentes selecionados.

Para o desenvolvimento do modelo de classificação do cancro de próstata, a arquitetura *Inception v3* foi implementada e treinada utilizando as subimagens da base de dados. Este modelo foi capaz de identificar três classes distintas: negativa (tecido benigno), positiva (tecido maligno) e neoplasia intraepitelial da próstata, esta última, embora com menor precisão dada a quantidade reduzida de exemplos pertencentes a esta classe. Um valor de 93 % de precisão foi obtido, o que corresponde a valores equiparados ao estado da arte para este tipo de técnicas. Este valor, contudo, demonstra ainda potencial para otimização e melhoria, uma vez que as diferentes classes dos dados utilizados seguiam uma distribuição não equilibrada. A inclusão de mais casos clínicos e a aplicação de técnicas de aumento de dados, podem ser facilmente realizadas, o que culminará num modelo com ainda melhor precisão de classificação.

Relativamente ao modelo referente à previsão de recidiva bioquímica, a mesma arquitetura foi utilizada, mas neste caso, treinada apenas com base nas subimagens positivas, isto é, as subimagens contendo tecido maligno da próstata. Os resultados obtidos revelaram que este modelo não tem a capacidade de extrair informação relevante correlacionada com o objetivo do estudo, e portanto, não consegue distinguir com sucesso casos não recorrentes de casos recorrentes, produzindo apenas uma precisão de 60 %. Contudo, apesar do referido modelo falhar na execução do objetivo estipulado, é fundamental notar que a tarefa de predição de recidiva bioquímica é de complexidade elevada, não sendo possível aos patologistas, através da observação das imagens histológicas, retirar nenhuma conclusão que diretamente se correlacione com esta condição. Diferentes abordagens, como por exemplo, o aumento da quantidade de dados utilizados, a introdução no modelo de características clínicas relevantes no prognóstico da doença poderão apresentar melhorias substanciais, no que diz respeito à capacidade preditiva deste modelo.

Concluindo, a capacidade de algoritmos de *deep learning* para extrair informação relevante de imagens digitais da histopatologia da próstata foi demonstrada através do presente estudo. O desenvolvimento e o criação de uma base de dados anotados, fornece a base fundamental para o desenvolvimento de modelos adicionais, onde diversas questões podem ser exploradas. O desenvolvimento de uma interface que permita implementar o modelo de deteção de cancro da próstata desenvolvido é também uma possibilidade, uma vez que fornece eficiência e consistência, beneficiando a prática da patologia clinica.

Palavras-Chave: Cancro da Próstata; Recidiva Bioquímica; Patologia Digital; *Deep Learning*.

Abstract

Prostate cancer is the second most prevalent cancer in men, worldwide. Histopathological assessment plays an indispensable role in understanding the disease mechanisms, providing definitive diagnosis to guide patient treatment and management decisions. The microscope has been the primary tool to this end, producing images at increased resolution. However, with the development of the first automated high resolution whole-slide imaging system, which allows the digitisation of glass slides, interest in using this system for different applications in pathology practice has steadily grown, giving rise to the digital pathology field. The promise of digital pathology is not, however, the simple transfer of an image from a glass slide to a monitor, not even the flexibility of distribution and modification of the image, but instead the potential to enhance the pathologist's assessment with information and artificial intelligence that cannot be perceived by human examination.

With the advent of digital pathology and the recent expert-level accuracy achieved by machine learning based algorithms in medical image detection, classification and segmentation, new possibilities to develop automated image analysis methods arise. As far as prostate cancer is concerned, these models have been aimed at supporting pathologist's image descriptions such as staging and grading, being hence, focused on the Gleason grading system. In order to explore alternative problems, the present dissertation presents data collection, processing and annotation procedures, that allowed the creation of an annotated digital histology database of prostate resection cases. These data was used to develop deep learning models not only to classify prostate cancer, but also to predict prostate cancer biochemical recurrence. Inception v3 architecture was implemented and trained from scratch for the proposed assignments.

The prostate cancer classification model yielded an accuracy of 93%, being able to identify three distinct classes: negative (benign tissue), positive (malignant tissue) and high-grade prostate intraepithelium neoplasia, the latter, although, with lower precision, given to unbalanced class distributions. The prostate cancer biochemical prediction model was not able to successfully distinguish between non-recurrent and recurrent cases, yielding an accuracy of 60%. This value was accomplished, nevertheless, due to the fact that the model was classifying all entries as negative, and therefore, the value of accuracy corresponds to the percentage of negative cases present in the dataset.

Although not all models here developed achieved good results, the capacity of deep learning algorithms to harvest relevant features from prostate histopathology digital images has been demonstrated. The development and establishment of an annotated database provides the fundamental basis to further develop additional models, and mainly to improve the biochemical recurrence prediction model by applying more sophisticated methods, given the complexity of this problem.

Keywords: Prostate Cancer; Biochemical Recurrence; Digital Pathology; Deep Learning.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN ..... **A**rtificial **N**eural **N**etworks

AUC ..... **A**rea **U**nder the **C**urve

BCR ..... **B**iochemical **R**ecurrence

BPH ..... **B**enign **P**rostatic **H**yperplasia

CAD ..... **C**omputer **A**ided **D**iagnosis

CNN ..... **C**onvolutional **N**eural **N**etworks

CT ..... **C**omputed **T**omography

DRE ..... **D**igital **R**ectal **E**xamination

FDA ..... **F**ood and **D**rug **A**dministration

H&E ..... **H**ematoxylin and **E**osin

MRI ..... **M**agnetic **R**esonance **I**maging

PCa ..... **P**rostate **Ca**ncer

PIN ..... **P**rostate **I**ntraepithelium **N**eoplasia

PSA ..... **P**rostate **S**pecific **A**ntigen

ReLU ..... **Re**ctified **L**inear **U**nit

RGB ..... **R**ed **G**reen **B**lue

SGD ..... **S**tochastic **G**radient **D**escent

SPCN ..... **S**tructure **P**reserving **C**olor **N**ormalization

SVI ..... **S**eminal **V**esicle **I**nvolvement

TCGA ..... **T**he **C**ancer **G**enome **A**tlas

TRUS ..... **T**rans**R**ectal **U**ltra**s**ound

WHO ..... **W**orld **H**ealth **O**rganization

WSI ..... **W**hole **S**lide **I**maging

1 INTRODUCTION

In this dissertation an approach to classification of prostate cancer and prediction of relapse is presented. The first chapter not only offers a brief contextualisation of why prostate cancer is a health issue, but also presents the goals of the project. Finally, the dissertation outline is also also described.

1.1 Context and Motivation

Men who are diagnosed with prostate cancer, and the physicians who care for them are faced with a wealth of information about this disease, including a variety of treatment options. While some cancers lay dormant for many years, others progress rapidly, being therefore, risk assessment a critical step in disease management, determining whether the cancer can simply be monitored, or a radical intervention by prostatectomy is needed. Among the involved procedures, the most important one is the examination of the biopsy and/or surgical tissue specimen by a pathologist. The image analysis of these prostate histopathological samples is used to provide the final detailed diagnosis and prognosis.

Current histopathology practice has several constraints, as it is highly time-consuming and it is often prone to inconsistencies and low agreement between pathologists. The staging and grading of prostate cancer is getting increasingly complex due to cancer incidence and patient-specific treatment options. The detailed analysis of a single case could require several slides with multiple stainings. Moreover, quantitative parameters are increasingly required (such as mitoses counting), increasing the pressure on pathologists to handle large volumes of cases, while still providing large amounts of information in the pathology reports. In this sense, the interest in digital pathology has continuously grown.

A high resolution automated system of whole slide imaging has been a strong advance for pathology, overcoming limitations, such as poor image quality and image navigation. It complements optical microscopies, also allowing for the application of visualisation software. This evolution can lead to virtual storage of images of pathological anatomy, and the associated establishment of global databases of histopathology images, which could be used both for clinical and educational purposes. On the other hand, it provides the opportunity to share clinical cases among health professionals, a key to clinical efficiency, given that the temporal window in which the diagnosis takes place is crucial for the hope of survival of many patients.

Nevertheless, the advantages of digital pathology are not yet sufficiently explored. There is a need to develop tools for automated tissue images processing and automated algorithm analysis

which can positively aid pathologists in decision making and improve the diagnosis/prognosis output. Recently, deep convolutional neural networks have revolutionised many medical image analysis domains, one of them being histopathology image grading and classification. Further exploration of deep learning approaches to histopathology tasks will certainly reveal its tremendous potential. Therefore, in this dissertation an approach to classification of prostate cancer and prediction of relapse is presented.

1.2 Objectives

The overall aim of the present work was to evaluate the power of machine learning based algorithms by developing and evaluating new classification and prognostic models based on histological digital images of the prostate. The secondary goals included:

(i) Comprehension of the main issues in classification and prediction tasks of pathology slides.

(ii) Analysis of the current literature in order to identify the state of the art workflow procedures and possible optimisation strategies for the problem.

(iii) Set up a database of histopathology digital images, as well as, respective annotations.

(iv) Implementation of the identified solutions to the developed database.

(v) Evaluation of the performance of the implemented algorithms and the respective clinical correlation.

1.3 Outline

In order to achieve the desired project goals the dissertation is divided into 8 Chapters. In each chapter, information considered relevant for the understanding of prostate cancer biochemical recurrence prediction problem and for the achievement of the proposed goals was presented. The current chapter introduces the context, motivation and general organisation of this written work.

In order to discuss the prostate cancer biochemical recurrence problem, a literature review is presented in **Chapter 2**. A comprehensive explanation of the relevant aspects surrounding not only the prostate gland, its anatomy and normal histology, but also prostate cancer and its diagnosis and prognosis is provided in **Section 2.1**. The role of digital pathology and its contribution to the current state of the art diagnosis and prognosis tools, is addressed in **Section 2.2**. In **Section 2.3**, the fundamental concepts regarding deep learning, neural networks and convolutional neural networks, a specific type of neural networks, is provided. **Chapter 3** presents a description of how the addressed problem is approached and which materials and methods are used to solve it. The results obtained from the developed work are detailed in **Chapter 4**, whereas in **Chapter 5** the respective results discussion is given. Finally, **Chapter 6** reports a brief overview of the main findings, the overall conclusions of the dissertation, as well as future perspectives.

2 LITERATURE REVIEW

This dissertation focus on different topics, namely prostate cancer, anatomic pathology and deep learning. Therefore, the presented chapter is devoted to present all of these concepts.

2.1 Prostate Gland

This section is dedicated to detail important aspects related to the prostate. An introduction of the normal prostate anatomy and histology will be presented and the remaining content is focused on prostate cancer: diagnosis strategies, histopathology, treatment options and relapse.

2.1.1 Prostate Anatomy and Histology

The human prostate is a walnut-sized organ of the male reproductive system, located in the pelvic cavity. Its function is to secrete fluid, which is one of the components of the semen. Anatomically, the prostate surrounds the urethra and its base is located at the bladder neck, while its apex at the urogenital diaphragm [1] (Figure 2.1). The urethra courses through the gland (known here as the prostatic urethra) and exits inferiorly at the apex. The Denonvilliers' fascia, a thin, filmy layer of connective tissue, separates the prostate and seminal vesicles from the rectum posteriorly [2].

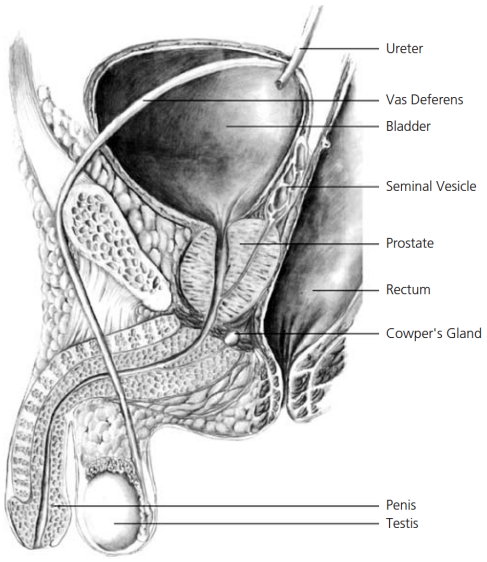


Figure 2.1: Sagittal section of the male pelvis. Adapted from [3].

In the twentieth century, the prostate gland was described, by several investigators, as being a lobular organ [4–6]. However, subsequent exploration of its anatomy established the current and most widely accepted concept of various prostate zones, rather than prostate lobes [7]. A three-zone model defines the transitional zone, central zone and peripheral zone (Figure 2.2). Additionally, there is the anterior fibromuscular stroma.

The transition zone consists of two equal portions of glandular tissue lateral to the urethra in the midgland. The central zone is a cone-shaped area of the adult gland, with the apex of the cone at the confluence of the ejaculatory ducts and the prostatic urethra. The peripheral zone comprises the prostatic glandular tissue at the apex, as well as, all the tissue located posteriorly near the capsule. This zone is commonly more prone to be affected by carcinoma, chronic prostatitis, and postinflammatory atrophy when compared to other zones. The anterior fibromuscular stroma forms the convexity of the anterior external surface, extending from the bladder neck to the apex of the prostate [8, 9].

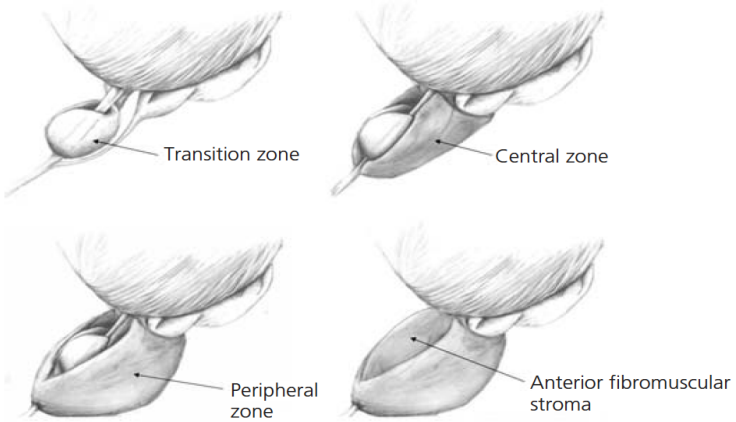


Figure 2.2: Zonal anatomy of the normal prostate as described by McNeal [7–9]. The transition zone comprises only 5%–10% of the glandular tissue in the young male. The central zone forms part of the base of the prostate and it is traversed by the ejaculatory ducts. The prostate is constituted by the peripheral zone, particularly distal to the verumontanum. Adapted from [3].

The prostatic capsule is composed of fibromuscular stroma which disappears towards the apex of the gland. Although the term "capsule" is commonly used and found in the literature, there is no consensus about the presence of a true capsule [10, 11].

The seminal vesicles are superiorly and posteriorly located to the base of the prostate. They undergo confluence with the vas deferens on each side to form the ejaculatory duct complex, which consists of two ejaculatory ducts along with a second loose stroma rich in vascular spaces. The seminal vesicles are resistant to nearly all the disease processes that affect the prostate. Seminal Vesicle Involvement (SVI) by prostate cancer is one of the most important predictors for cancer progression [12, 13].

As far as innervation is concerned, the prostate is an extremely well-innervated organ. Two neurovascular bundles are located posterolaterally adjacent to the gland and form the superior and inferior pedicles on each side. Different studies [14, 15] have, previously, demonstrated the importance of these nerves in penile erection, hence urologists as well as patients have put an increasing interest in nerve-sparing surgical treatment of PCa.

Histologically, the prostate is a branching duct-acinar embedded in a fibromuscular stroma. The epithelium has two layers: the luminal/secretory and the basal layer with some neuroen-

docrine cells in the epithelium. The intraluminal contents of otherwise normal prostatic glands include degenerate epithelium cells, corpora amylacea and calculi [11]. The prostate stroma includes fibroblasts, smooth muscle, vasculature and nerves and assorted infiltrating cells (e.g., mast cells, lymphocytes).

2.1.2 Prostate Cancer

According to the World Health Organization (WHO), cancer is a leading cause of death worldwide, accounting for 18.1 million new cases and 9.6 million deaths in 2018. Prostate Cancer (PCa) is one of the most common and dreadful types of cancer. Its incidence accounts for 1.3 million of the total cases (7.1%), ranked as the second most frequent form of cancer in males, following lung cancer. Annually, an estimated 350,000 men die with PCa [16, 17]. These global patterns are also backtracked in Portugal, where PCa incidence rate is around 121 per 100,000 inhabitants, the highest cancer incidence rate, both in men and in total. Of these, 22% died in 2010 [18, 19]. As a result, PCa, as many other types of cancer, is, undoubtedly, one of the major societal challenges in healthcare, particularly among men.

PCa growth can be characterised by two main types of evolution: slow-growing or clinical insignificant tumours and clinical significant tumours. The former comprises up to 85% of all prostate cancers, where the tumour is usually confined to the prostate gland. For these cancers, the treatment is replaced with frequent monitoring [20]. The latter type of PCa - the clinical significant tumours - progresses rapidly, metastasising from the prostate gland to other organs and bone. Hence, radical surgery and/or radiation therapy are the treatment approaches used to eradicate this type of tumour [21, 22].

2.1.2.1 Prostate Cancer Diagnosis

PCa has no specific presenting symptoms and is frequently clinically silent [23, 24]. Therefore, in the past decades, the diagnostic value of screening by using different detection strategies has come to the forefront of focus. With the introduction and widespread of screening with serum Prostate-Specific Antigen (PSA) blood test, the number of men diagnosed with PCa has increased, with most cases being at initial phases - localised disease - instead of advanced stages of cancer.

The initial detection of PCa often occurs during routine medical exams. The physician has two screening options: a PSA test and a Digital Rectal Examination (DRE). A PSA test is a blood test aimed to measure the PSA protein level in the patient blood (in healthy men, the PSA level is mostly below 4 ng/mL (nanograms per milliliter)). However, PSA is produced by both normal and malignant prostatic epithelial cells, making it an organ-, but not cancer-specific biomarker, since benign conditions - such as prostatic infection, and benign prostatic hyperplasia [25] - may also cause its level to rise. Hence, DRE allows the physician to discern PSA clinical significance, by detecting palpable abnormalities (e.g., nodules, asymmetry, hardness) in the posterior and lateral regions of the prostate gland, where the majority of cancers develop [26].

Further verification of suspicious clinical or biochemical tests is based on histopathologic verification of a prostate biopsy [27]. During a transrectal ultra-sonography-guided biopsy - which is the currently most commonly used technique - multiple samples of prostate tissue are collected. By extracting more samples, the higher cancer variability can be identified, avoiding not only false negatives, but mainly distinguishing indolent prostate cancers with low risk of

progression – which only require surveillance – from clinically significant cancers – which require a radical therapy.

To perform the definitive diagnosis, the collected samples are observed and examined under a microscope by a pathologist, who evaluates different histologic characteristics to determine whether PCa is present.

2.1.2.2 Pathology of the Prostate Cancer

Most prostate cancers are of prostatic glandular origin, i.e., adenocarcinoma, though occasionally, other forms of cancer, such as transitional cell carcinoma, squamous cell carcinoma and undifferentiated carcinoma, may arise within the prostate [3]. The diagnosis of prostate carcinoma relies on a combination of architectural and cytological features. The architectural features are assessed at low to medium power magnification, with emphasis in spacing, size and shape of acini. Malignant acini usually have an irregular arrangement, as well as, an irregular contour. The groups of acini are either closely packed in back-to-back arrangement without intervening stroma or are haphazardly distributed. In addition, while in benign acini an intact basal cell layer is present, in prostate cancer, only a single layer of epithelium cells is present. Therefore, the basal cell layer is a crucial feature to assess, in order to differentiate between benign and malignant cells. Regarding the cytological features, nuclear enlargement and hyperchromasia, as well as, mitotic features represent the main characteristics of malignant tumour focus [3, 11].

However, most of these features, when confined, are not entirely specific to malignancy [28]. A pseudo-infiltrative pattern, for instance, can also be found in benign conditions, such as atrophy, while nuclear abnormalities might be found in benign glands adjacent to inflammation areas [3]. Additionally, biopsies are regarded as difficult to resolve, not only due to the non-targeted nature of the procedure, but also because only a few cores are available for diagnosis, where a small number of atypical cells are not sufficient to allow an unequivocal diagnosis [29–31].

Therefore, a number of other lesions and conditions (Figure 2.3) associated to the prostate must be considered within the context of cancer diagnosis, since they are *mimickers* of prostate adenocarcinoma [28, 32, 33]. Of these, Benign Prostatic Hyperplasia (BPH), atrophy, and Prostatic Intraepithelial Neoplasia (PIN) must be highlighted.

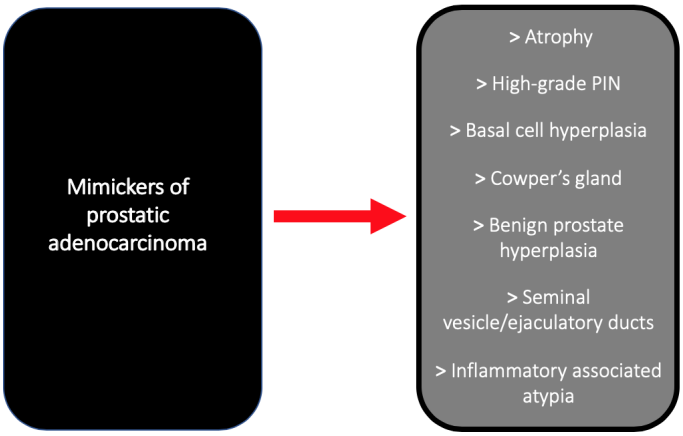
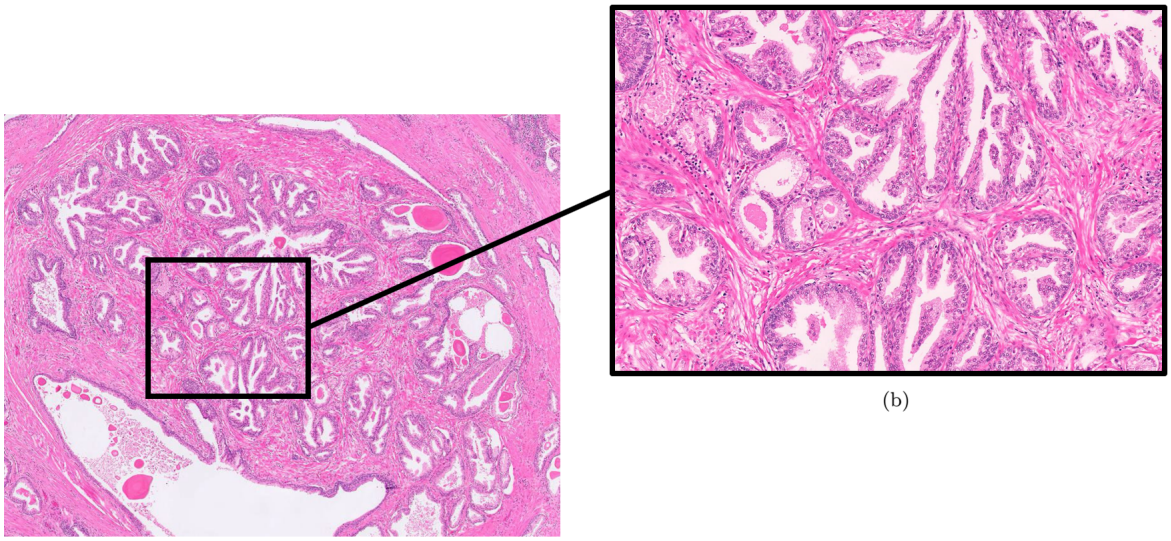
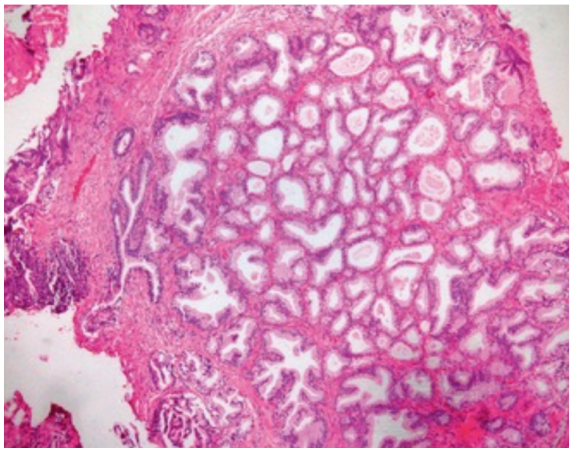


Figure 2.3: Lesions and structures that can simulate prostate cancer. Some of these lesions are common differential diagnoses of prostate adenocarcinoma.

BPH (shown in Figure 2.4) is one of the most common diseases in men from 40 years onwards, probably due to the alterations in hormone balance. It produces obstruction to bladder outflow and as a result of pressure effects on the proximal conducting system of the urinary tract [11]. Microscopically, BPH is characterised by the development of a number of nodules, which can have different origins - stromal, glandular or mixtures thereof. Nodules formation is expected to be detected more often in specimens from prostate excision, instead of from biopsies. In the latter, these nodules are sometimes misdiagnosed as adenocarcinoma [33, 34].



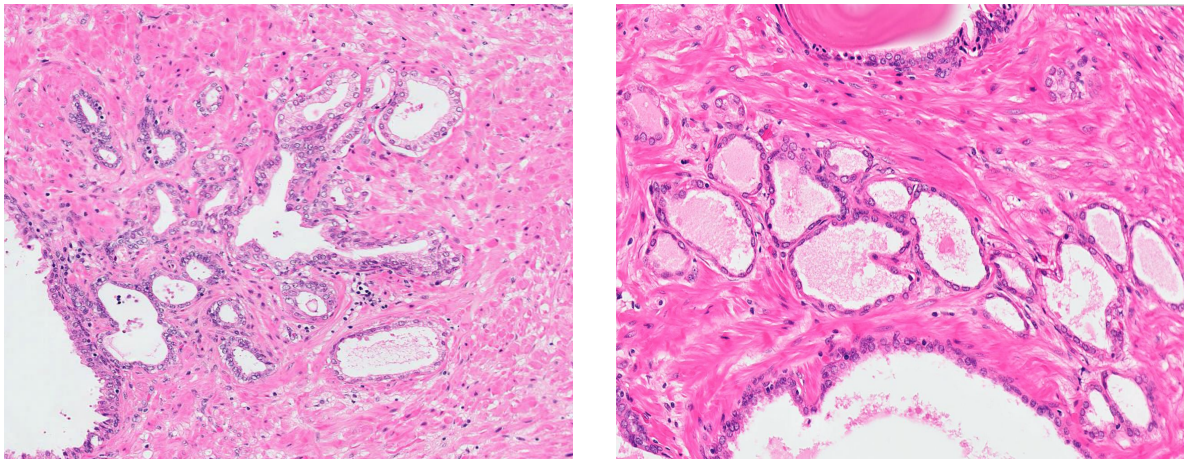
(a) (b)



(c)

Figure 2.4: BPH foci in (a) low magnification and in (b) in high magnification. The intraluminal infoldings within the glands is charcteristic of BPH. (c) Epithelium predominant round nodule is typical in BPH. A small sample of this nodule discovered in a biopsy could be misdiagnosed as prostate adenocarcinoma.

Atrophy - Figure 2.5 - is a common age-related benign condition, most frequently seen in peripheral zones samples. Any subtype of atrophy is expected to show basal cells, however their detection in some biopsies may be difficult [35] - Figure 2.6.



(a) (b)

Figure 2.5: Low magnification of different patterns of atrophy. (a) Acinar atrophy. (b) Cystic atrophy.

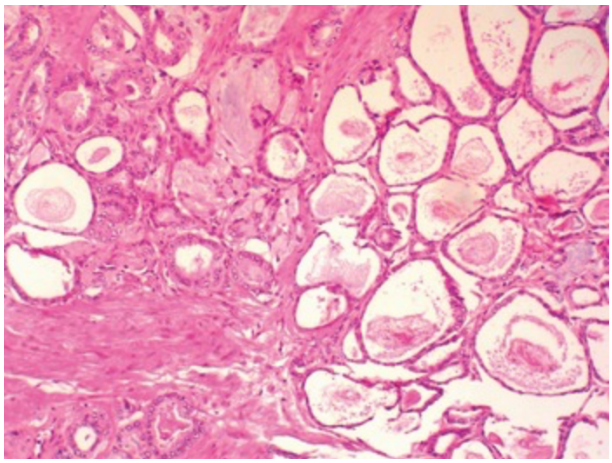
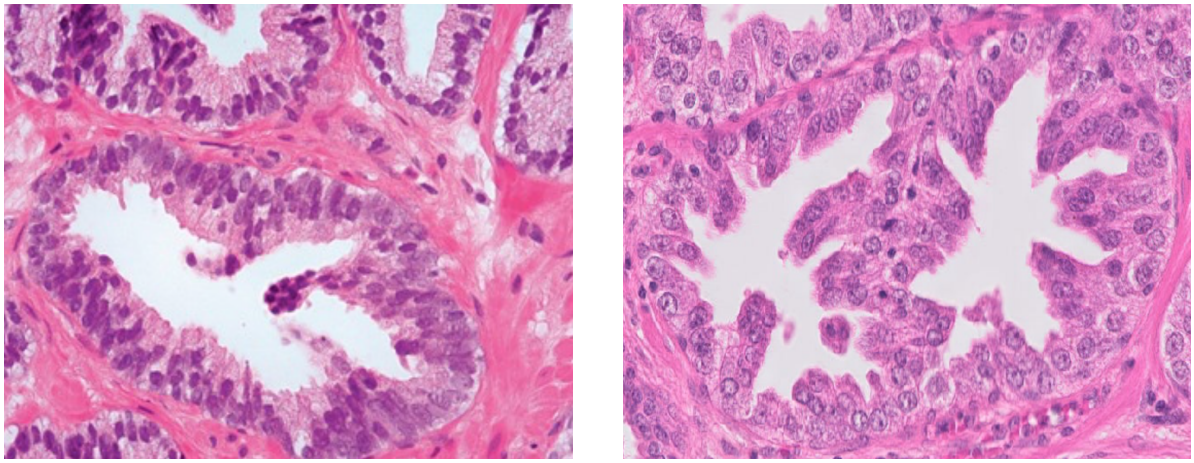


Figure 2.6: Atrophic prostate carcinoma may be occasionally cystic and resembling a focus of benign atrophy.

PIN refers to multiple foci of cytologically atypical luminal cells overlying diminished number of basal cells in prostatic ducts and is a forerunner of invasive prostatic carcinoma [11]. According to the cytologic characteristics of its cells, PIN may be classified into low grade PIN and high-grade PIN (Figure 2.7). The former is composed of enlarged cell nuclei, which vary in size and have a slightly increased chromatin content. The latter is characterised by cells with large nuclei of uniform size and prominent nucleoli - similar to those of carcinoma cells . The basal cell layer is intact on lower grade PIN, but may present disruption in higher grade PIN [36]. PIN can be erroneously evaluated as carcinoma or *vice-versa* [28], especially because there are a specific type of carcinoma which morphologically resemble PIN, the PIN-like ductal carcinoma Figure 2.8. The four different types of high-grade PIN are represent in Figure 2.9: tufting, micropapillary, cribriform and flat.

When several of the above-mentioned cancer features are detected in a specimen, the diagnosis of prostatic adenocarcinoma can be confidently performed and a classification prognostic grade - the Gleason Score - is reported. The Gleason score system, developed in the 1960s to determine the aggressiveness of the cancer, is based on glandular architecture, which can be divided into five patterns of growth or grades. Pattern 1 lesions consist of nodules of small, well-defined glands with limited infiltration of the surrounding tissue. In contrast, pattern five



(a) Low-grade PIN. **(b)** High-grade PIN.

Figure 2.7: High power view of the different PIN categories: low grade and high-grade.

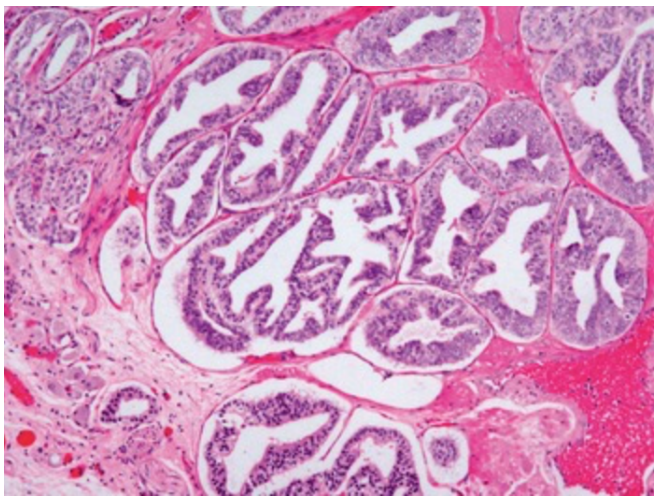


Figure 2.8: Example of carcinoma of the prostate, which resembles a PIN foci. It is denominated PIN-like ductal carcinoma.

lesions consist of layers of malignant cells with no discernible glandular differentiation and which are widely infiltrative. Due to the fact that most tumours are highly variable, including two or more components of these patterns, the Gleason score is, therefore, obtained by adding two sub-grades: primary and secondary grades [37]. The primary grade is assigned to the dominant pattern of the tumour, while the secondary is assigned to the subordinate pattern - for instance Gleason score 4+3.

Although changes to this grading system have been implemented to this grading system, with complementary group grades being adopted, the Gleason score system remains as one of the most significant factors in the clinical decision-making process, not only in biopsy specimens, but also mainly after a radical intervention is performed, being a predominant tool in outcome prediction [38].

2.1.2.3 Prostate Cancer Treatment

After combining the results of the PSA, DRE and histological grade of the biopsy specimen, the stage and nature of the cancer is determined. If the tumour has metastasized beyond the

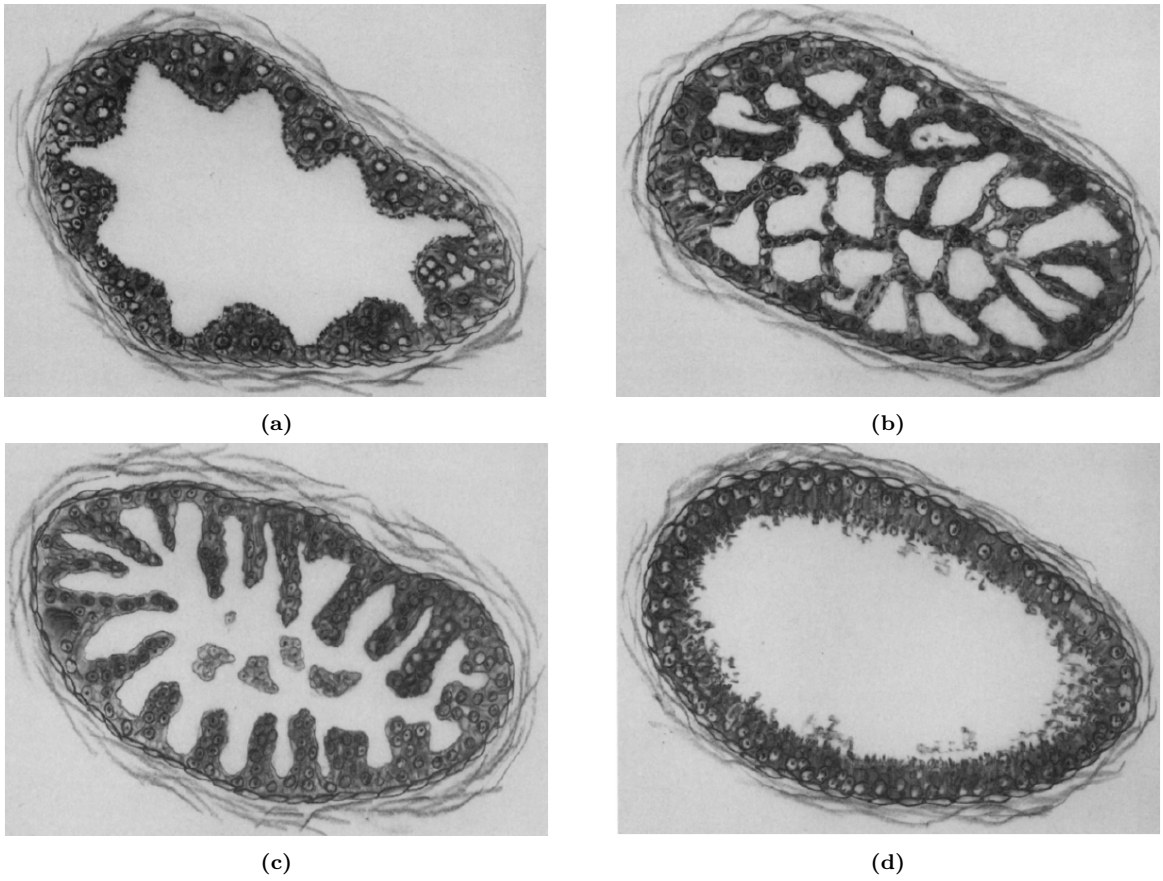


Figure 2.9: Schematic illustration of architectural patterns of high-grade PIN. (a) The tufting pattern shows undulating mounds and heaps of cells protruding into the lumen. (b) The micropapillary pattern shows delicate finger-like projections, often with bulbous tips, with or without flbrovascular cores. (c) The cribriform pattern shows a sieve-like pattern. (d) The flat pattern shows one or two layers of cells. Adapted from [39]

prostate gland, additional tests might have to be performed, including Magnetic Resonance Imaging (MRI), bone scan, Computed Tomography (CT) and TransRectal Ultrasound (TRUS). These techniques provide relevant information, as far as accurate staging is concerned, both in local and advanced disease. If a patient is found to have advanced disease with extended involvement of surrounding structures - SVI, perineural and lymphovascular invasion - a radical treatment would be an inappropriate approach to employ, since it would unnecessarily endanger the patient while burdening healthcare resources [3]. In these cases, radiation therapy, androgen-deprivation treatments and chemotherapy are the therapies of choice. Nevertheless, advanced prostate cancer remains an incurable disease and new alternative treatments are required.

Conversely, when the diagnosis reveals localised PCa, a prostate confined tumour with no extension to the seminal vesicles, regional lymph nodes, or distant site, he is a candidate for surgical prostatectomy. This procedure is the standard first-line treatment with curative intent [27, 40]. The perceived advantage of radical prostatectomy is that, the entire prostate, where the tumour is confined, is surgically removed, aiding towards the reduction of the metastasis risk [41]. Moreover, with the recent introduction of robotics in minimally invasive surgery, enhanced treatment performance has been reported. Robotic assisted laparoscopic radical prostatectomy has gained acceptance among urologists and patients, not only because it allows the surgeon to perform complex surgical procedures with both dexterity and minimal fatigue, but mainly because it is a less invasive procedure, diminishing operative time, blood loss and hospitalisation

time. Suggestion of earlier improved erectile function and full sexual health, after robotic surgery have also been reported [42, 43].

However, despite the improved results of these robotic surgeries, cancer recurrence is a concern for men undergoing definitive treatment with curative intent for localised PCa [44]. Of all patients treated with radical prostatectomy, approximately 35% are reported to experience a re-emergence of PSA serum, after its total abrogation: a state known as BioChemical Recurrence (BCR) [45, 46]. A detectable PSA level implies residual prostate tissue and, most likely, recurrent PCa.

Considering the above, the history of BCR following surgery is highly variable, since PSA relapse has different meanings according to clinicopathological features - such as: Gleason Score, clinical stage and surgical margins status [47]. Therefore, it is of paramount importance to treating physicians, to assess both, 1) whether BCR is going to emerge after surgery, and 2) distinguish different biochemical recurrence based on preoperative and/or postoperative parameters. The outcome prediction information will aid towards clinical decision-making, by implementing the golden complementary therapies, achieving, thus, a more personalised and cost-effective treatment regimen.

2.1.3 Prostate Cancer Biochemical Recurrence Prediction

A method capable of forecasting recurrence is highly desirable, due to the complexity of PCa disease and the associated decision-making process. Therefore, accurate estimates of the likelihood of cancer diagnosis, stage, clinical significance, treatment success, complications, and long-term morbidity are essential for patient counselling and informed decision-making. By understanding the most probable endpoint of a patient's clinical course, physicians may modify treatment and post-treatment strategies, in order to balance benefits and adverse effects of a specific therapy. Prediction also allows patients to choose responsibly among the different treatment strategies, proposed by the clinicians, which, ultimately, may improve their satisfaction after treatment [48].

Traditionally, physician judgement is the basis for risk estimation, patient guidance, and decision-making. However, clinicians' estimates are often biased due to both subjective and objective confounders [49–51]. To overcome these limitations and to obtain more accurate predictions, researchers have developed predictive tools – directed at predicting the probability of an outcome of interest for the individual patient, without considering the effect of time [52].

Currently, clinical prognosis of PCa relies on the evaluation of histology sections under a microscope, which allows the pathologist to determine specific tumour characteristics: Gleason score, capsule invasion, extra capsular extension, surgical margins, SVI, lymph node invasion. These features serve as variables in the widely-validated Kattan nomogram [53, 54] and CAPRA-S score [55] predictive methods, considered as the best performing tools for prostate outcome prediction after radical prostatectomy. Although these nomograms remain the most widely used tools in clinic, as far as postoperative clinical management is concerned [41], their modest overall accuracy is further devalued by the realisation that these tools largely fail in the most common cohort in the modern era (mid-grade, confined disease with moderate PSA level) [56, 57]. A tendency to overestimate the likelihood of disease recurrence for lower risk patients [55, 57] has also been reported. Hence, and in addition to these obvious generalisation and complexity limitations [58, 59], the lack of periodic updates to reflect changes in patient and

disease characteristics may compromise the applicability of some models to contemporary patient populations. To illustrate this point, the surgical procedure currently employed is Robotic-Assisted Radical Prostatectomy (RARP), as opposed to the invasive laparoscopic or open radical Prostatectomy previously adopted (until around 2010) [60, 61]. This consideration is likely to influence the outcome prediction.

While most prognostic efforts have focused on incorporating additional clinical variables, which has not significantly improved predictions [62], the potential of digital histologic images has not, yet, been fully explored. An unanswered question is whether these tools for outcome prediction could be bypassed by novel machine learning techniques, with the power to directly learn the prognostically relevant features in microscopy digital images of the tumour, without prior identification of the known tissue entities [63], e.g. Gleason score, mitoses, infiltrating immune cells, tumour budding.

Therefore, in this dissertation, and in particular in Section 2.2, emerging techniques such as digital pathology and deep learning will be exploited in order to overcome the shortcomings observed in the tools currently used to predict BCR. While digital pathology allows digitisation of the histopathology image and its consequent quantitative analysis, deep learning has the power to enhance information that cannot be perceived by the human eye. It is expected that a model resulting from the combination of these techniques has the power to predict the occurrence of relapse in patients treated with radical prostatectomy.

2.2 Anatomic Pathology

In this section the field of anatomic pathology will be detailed. In first place, the fundamentals of anatomic pathology will be explained. Then, the role of an emerging technique as digital pathology will be discussed and their advantages and limitations presented. Finally, approaches using deep learning applied to histopatholgy in precision medicine will be presented and some image classification examples for prostate cancer will be detailed.

2.2.1 Fundamentals of Anatomic Pathology

Interpreting images of tissues and cells at a resolution higher than the naked human eye is the core of pathology (also referred to as histopathology or anatomical pathology). This medical specialty is fundamental to studying, understanding and diagnosing disease, while providing further insight into what is required for a successful treatment.

As disease processes occur at the molecular/cellular levels, its manifestations and mechanisms can be observed at the tissue level using a microscope. In order to do so, several steps from the acquisition of the sample to the final diagnosis, must be completed (Figure 2.10).

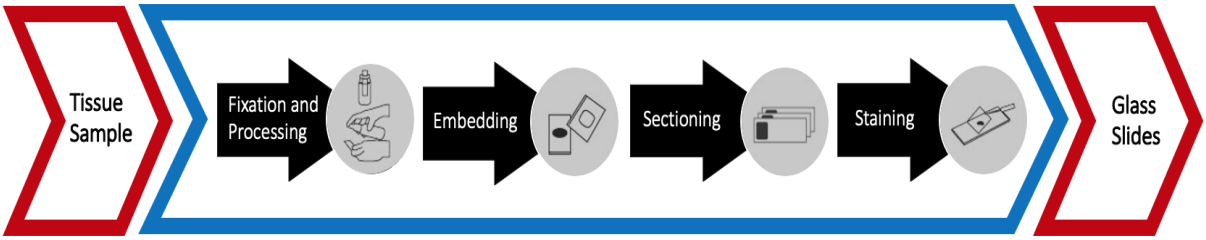


Figure 2.10: Diagram representing the histology process which allows a diagnosis to be rendered.

The clinical histology process begins when the portion of tissue to be studied is extracted from the patient. Several possible approaches to collect tissue are possible, including fine-needle aspiration, needle biopsy, excisional biopsy or excision of the entirety of the lesion. The sensitivity (i.e. the likelihood of getting the correct diagnosis) and specificity (i.e. the likelihood of not getting the incorrect diagnosis) increase from fine-needle aspiration to excision of the entire lesion, since the larger the biopsies, the more cellular content is preserved. Hence, the pathologist is able to examine multiple slides from different regions of the sample. In order to stop tissue degradation and prevent microorganisms growth, fixation of the sample, usually accomplished with formalin, must be performed [64].

Afterwards, the tissue is physically and chemically stabilized, during the processing phase, where water intrinsically contained within the tissue is removed, and a hardening agent (paraffin) is infiltrated within the tissue, allowing the sample to become firm. During the embedding phase, the sample is inserted into a block of support rigid mould. The result is a *tissue block* which can now be sliced into thin sections (3-10 μm thickness). At this point, the tissue slices are nearly invisible under a light microscope and must be stained to create contrast. The most widely used staining method, both for diagnostic and research purposes, is Hematoxylin and Eosin (H&E). H&E histological staining protocol differentiates structures according to their pH. Specifically, hematoxylin, of a basic nature has affinity for acid/basophilic substances, such as nucleic acids (DNA, RNA). In contrast, eosin has affinity for basic/acidophilic structures, of which cytoplasm is an example. Hence, when a biological sample stained with H&E is observed under a microscope, the cell nuclei have a blue-purple hue, whereas cytoplasms have shades in a range from pink to red. The final result, following all of these processes, is a slide or set of slides prepared to be visualised and analysed under a microscope by a pathologist [64–66].

2.2.2 Digital Pathology

Since the 19$^{\text{th}}$ century, pathologists have been using the microscope as the primary tool to observe stained specimen on the slide glasses. They rely on the microscope, since it has the power to provide live images at increased resolution [67]. However, after the development of the first automated, high resolution Whole-Slide Imaging (WSI) system by Wetzel and Gilbertson in 1999, interest in using WSI for different applications in pathology practice has steadily grown [68, 69]. This process was catalysed by changes in imaging hardware, as well as, growth in computational power accompanied by faster and increased data storage and declining computational costs [70].

The integration of digital technologies systems in anatomic pathology departments offers benefits, not only in terms of quality, but also efficiency [71]. WSI – the digitisation of glass slides – produces high resolution digital images, involving relatively high speed digitisation of glass slides of different samples (e.g. tissue sections, smears), scanning them at multiple magnifications and focal planes (x, y and z axes) [72]. Essentially, a whole slide image consists of a pyramid of tiled images, with various levels corresponding to a specific magnification of the image - Figure 2.11.

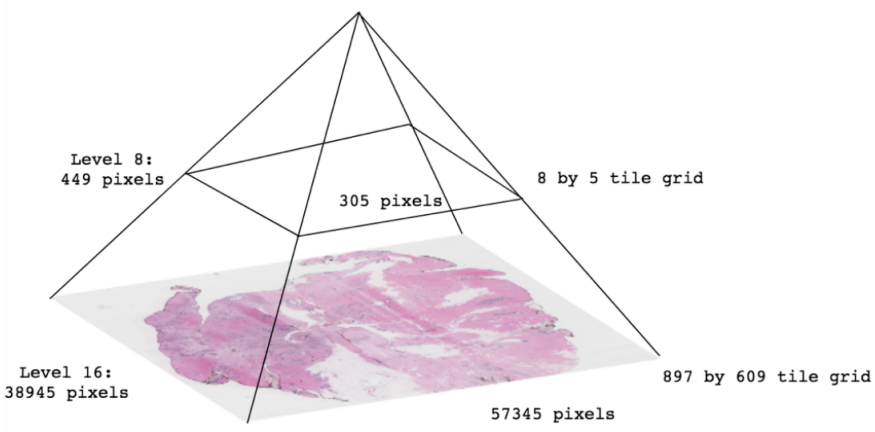


Figure 2.11: Pyramid structure representing the various resolutions that compose a whole slide image. Adapted from [73].

Interactive software for visualisation enables the slides to be examined on a computer screen, allowing the pathologist to freely navigate the image over a complete range of standard and in-between magnifications and perform, therefore, tasks that have historically been carried out with a light microscope. Digital platforms are very useful for sharing cases both locally and internationally and to provide expertise to remote locations. Consultation for difficult cases is, also, possible within hours instead of days to weeks for cases sent through regular mail. In addition to this, providing a digital record of the exam to the patient, while maintaining the original physical glass slides in archive is a tremendous advantage, as far as clinical management is concerned. A typical digital pathology set-up is shown in Figure 2.12.

Digital slides, in contrast to glass slides, can be easily annotated and regions of interest identified and associated to the written report. Instead of just replacing the microscope for diagnostic purposes, digital pathology provides improved quantification analysis of histopathological images, which are tasks highly affected by subjectivity. These image analysis techniques are more reproducible than traditional microscopy for quantitative measurements [74].

Recently, the US Food and Drug Administration (FDA) cleared the marketing of the first WSI system for digital pathology [75]. This opens a new era for the introduction of computation components into each diagnosis process. Moreover, the digitization of high-resolution whole slide images makes it possible to implement Computer Aided Diagnosis (CAD) system to analyze large-scale image data, thus alleviating intra- and inter-observation variations among pathological experts and achieving statistical conclusions across hundreds of slides [76]. Digital pathology is regarded as the required "bridge" to apply artificial intelligence algorithms to whole slide images which has the potential to greatly improve clinical practice.

2.2.2.1 *Intelligent* Digital Pathology

The promise of digital pathology is not, however, the simple transfer of an image from a glass slide to a monitor, not even the flexibility of distribution and modification of the image, but instead the potential to enhance the pathologist's assessment with information and intelligence that cannot be perceived by human examination. Moreover since most pathology diagnosis are, currently, based on the subjective (but educated) opinion of pathologists, there is, clearly, a necessity for quantitative image-based assessment of digital pathology slides.

Histopathology slides provide a comprehensive overview of disease and its effect on tissues,

Figure 2.12: Digital pathology workstation. The microscope has been replaced by a high resolution computer screen where the pathologist is able to view digitally scanned slides. Adapted from [74].

since the preparation process preserves the underlying tissue architecture [77]. As a result, a detailed spatial interrogation (e.g. capturing nuclear orientation, texture, shape, architecture) of the entire tumour morphological landscape and its most invasive elements can be deduced only from a histopathology image. The additional complexity and density in these images, while providing a wealth of information, is ideally suited to power automated computed algorithms. The application of these algorithms is essential, not only from a diagnostic perspective, in order to understand the underlying reasons for the rendering of a specific diagnosis (e.g. specific chromatin texture in the cancerous nuclei which may indicate certain genetic abnormalities), but also for research applications (e.g. to understand the biological mechanisms of the disease process and ultimately to improve diagnosis and treatment procedures). Consequently there has been substantial interest in the digital pathology image analysis community to develop algorithms and feature identification approaches for automated tissue classification, disease grading and also developing histologic image-based companion diagnostic tests for predicting disease outcome and precision medicine [78].

Recently, a breast cancer multi-classification method using a newly proposed deep learning model was reported [79]. Instead of just focusing on the binary classification (positive or negative for malignancy), these authors further explore quantification to obtain different tumour sub-classes within benign and malignant conditions (e.g. ductal carcinoma, lobular carcinoma, adenosis, fibroadenoma,...). The model used data from a large-scale dataset – BreaKHis [80] – achieving a remarkable performance (average 93.2% accuracy), when compared to previous reported techniques.

Coudray et al trained a deep convolutional neural network (Inception v3) to automatically analyse lung tumour digital slides, and distinguish between lung adenocarcinoma, squamous cell carcinoma or normal lung tissue. The deep learning model was developed using 1634 whole-slide images available in The Cancer Genome Atlas (TCGA) [81] and tested on independent cohorts in-house collected. Performance of the reported method was comparable to that of pathologists, with an average Area Under the Curve (AUC) of 0.97 [82].

Moreover, a predictive tool resulting from the combination of convolutional and recurrent architectures was recently developed to predict colorectal cancer outcome. This approach was

based on digitised tumour tissue samples from 420 patients, without any intermediate tissue classification. The results show that deep learning-based outcome prediction outperforms visual histological assessment performed by human experts in the stratification into low- and high-risk patients [83].

For automated Gleason PCa grading, *Arvanati et al* has recently reported a deep learning approach to analyse prostate digital samples. A convolutional neural network was trained using detailed Gleason annotations on a cohort of 641 patients and was then evaluated on an independent test cohort of 245 patients annotated by two pathologists. The model's Gleason score achieved pathology expert-level stratification of patients into prognostically distinct groups. Overall, this shows promising results regarding the applicability of deep learning-based solutions towards more objective and reproducible PCa grading, in contrast with a human subjective evaluation, especially for cases with heterogeneous Gleason patterns [84].

These and other recent studies [78, 85–89] suggest that state of the art deep learning techniques applied to digital slide images can improve diagnostic accuracy and efficiency. It can also provide features correlated with prognosis that a human observer cannot perceive. As these and more studies unfold, digital pathology will undoubtedly open up new avenues for computational exploration of individual disease tissues. With the emerging of deep learning tools, and many available pre-trained networks, the accurate prediction of a disease outcome – which is one of the most interesting and challenging tasks for physicians in the cancer medical field – will start to be accomplished. Ultimately, better disease surveillance, earlier detection, more accurate and consistent diagnosis and prognosis will create an era of truly personalised medicine.

2.3 Deep Learning

The present section is dedicated to providing insights about deep learning. Firstly the basic concepts of neural networks, the basis of deep learning, will be presented and the rest of the chapter is dedicated to CNN, a special type of neural networks, which has been successfully employed in different tasks, involving processing data with grid-bases structure, especially images.

2.3.1 Basic Concepts of Neural Networks

Deep learning offers a different perspective on feature learning and representations, where robust, abstract and invariant features are learnt end-to-end, hierarchically, from raw data [90]. Despite being a subfield of machine learning, the key difference is on how features are extracted. While traditional machine learning approaches depend on handcrafted engineering features, deep learning algorithms have the power to learn a set of optimal features automatically/user-independently. This means that these algorithms automatically perceive the relevant features required for the solution of the problem. Deep learning is inspired by artificial neural networks (ANN), - or simply neural networks - which in turn are biologically inspired algorithms. The term network is named after the topology, since neural networks are represented by composing many different layers together. A network composed of three layers connected together in a chain structure is represented in the following figure - Figure 2.13.

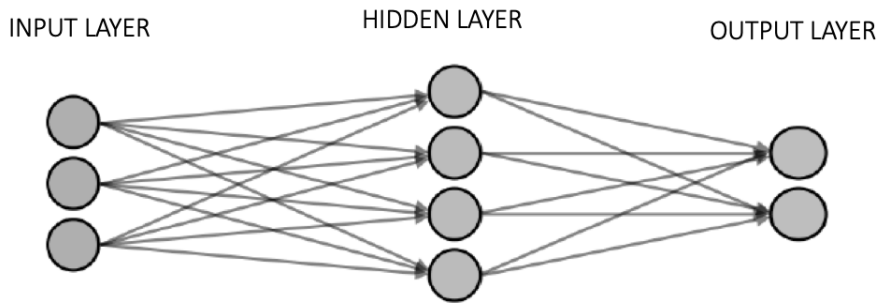


Figure 2.13: Architecture of simple neural network. It is composed of 3 layers: input layer, hidden layer and output layer. The input layer receives the input data, a vector of N elements, in this case N = 3. The hidden layer is composed of four units, thus the width of the model is four. The output layer computes the answer of the network from the hidden representation.

The first layer of a neural network is called the input layer, while the last layer is called the output layer. The former holds values of the input vector and the latter holds the final results provided by the model. The layers between the input and output are named hidden layers. The behavior of these layers is not directly specified by the training data i.e. the learning algorithm must decide how to use these layers to best implement an approximation to the desired output. The overall number of hidden layers indicates the depth of the model - thus the term deep -, while their dimensionality (i.e. number of neurons) determines the width of the model.

The fundamental component of neural networks is the artificial neuron, named node or unit. Just as in biological neurons, an artificial neuron takes in a vector of inputs $x = [x_1\ x_2\ x_3\ ...\ x_n]$ each of which is multiplied by a specific weight $w = [w_1\ w_2\ w_3\ ...\ w_n]$ and an activation function ϕ is applied. The output of the neuron is expressed as:

$$y = f(x) = \phi(\mathbf{x} \cdot \mathbf{w} + b) \tag{2.1}$$

where b is the bias term. Weights and biases are the neuron's parameters, and their values are determined in the process of learning. A schematic representation of an artificial neuron is shown in Figure 2.14. The activation function decides whether a neuron should be activated or not, i.e. whether the information that the neuron is receiving is relevant or whether it should be ignored. The purpose of the activation function is to introduce nonlinear properties, allowing the network to learn from highly complex functions, instead of just computing linear functions (e.g. the linear transformation of the input by the weights and biases).

As far as activation functions are concerned, the most commonly used types are Sigmoid, Tanh, Rectified Linear Unit (ReLU) and Leaky ReLU. However, ReLU is the function adopted in the majority of the deep learning models, since it is computationally economical compared to the remaining alternatives and has proven to successfully work in a wide range of different cases.

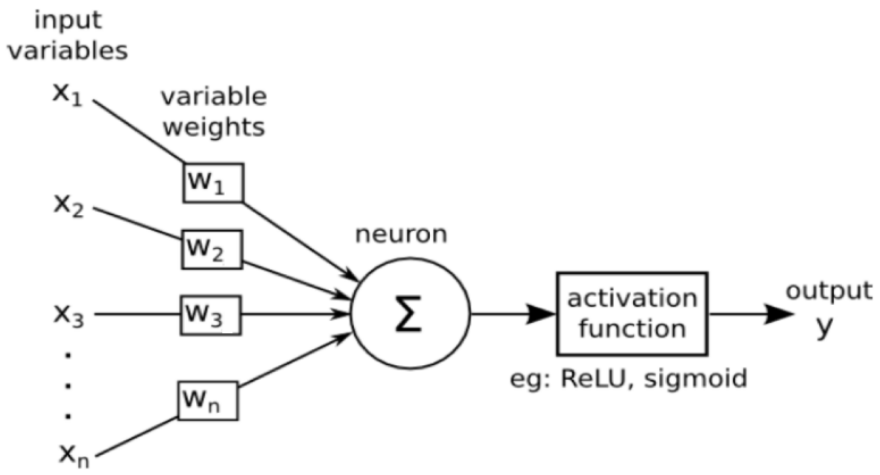


Figure 2.14: Schematic representation of an artificial neuron, where $\sum = \mathbf{x} \cdot \mathbf{w} + b$.

As far as the information flow is concerned, different models are distinguished in neural networks: feedforward and recurrent neural networks. In feedforward networks, information flows from the input to the output through the different network layers and without feedback connections. If those connections are present and intermediate outputs are redirected back to previous layers or stored for later use, the network is called recurrent. Since only feedforward networks are used in this study, the following sections concern only this type of networks.

2.3.2 Training a Neural Network

Training a neural network, i.e. learning the values of the parameters (weights and biases) is the essential part of deep learning and is accomplished by solving an optimisation problem. In order to solve this optimisation problem, gradient-based algorithms are used and the weights of the model are updated using the backpropagation algorithm [91].

The first phase is the forward-propagation, which occurs when the network is exposed to the training data and these cross the entire neural network in order to calculate the predictions (labels). In detail, the input data is passed through the network in such a way that all the neurons apply their transformation to the information they receive from the neurons of the previous layer and send it to the neurons of the next layer. When the data has crossed all the layers, and all its neurons have made their calculations, the final layer will be calculated with a result of label prediction for the original input samples.

Afterwards, a loss function - often called the objective function, cost function, or simply error - must be chosen. This function is used to evaluate to what extent the output is correctly predicted, i.e. it quantifies how good or bad the prediction result is, when compared to its known label. Ideally, the cost function would be zero, meaning that there is no divergence between the estimated value and the expected value. Typically, for classification problems the cross-entropy loss is the most commonly used loss function. Once the loss has been computed, this information is propagated backwards in order to improve the model predictions. Hence, the name of the algorithm: backpropagation. Starting from the output layer, that loss is propagated to calculate the gradient of the loss function with respect to each neuron of weights or parameters. Thereafter, these parameters - weights - are updated in the opposite direction of the gradient of the loss function. Different variants of gradient descent algorithm have been proposed, includ-

ing Stochastic Gradient Descent (SGD), Adaptive gradient (Adamgra) and Adaptive moment estimation (Adam). The Adam optimiser - derived from the Adamgrad - is very similar to the SGD, however, instead of maintaining a single learning rate for all weight updates it computes and uses individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients.

This process is then repeated for each example of the dataset and the weights of the interconnections of the neurons will be gradually adjusted until good predictions are obtained, i.e. until a minimal loss error is reached.

2.3.2.1 Hyperparameters Tuning

In addition to the weight parameters defined within the network, deep learning algorithms also require additional parameters to carry out the learning process. These hyperparameters highly influence the algorithm performance and must, therefore, be wisely selected.

In order to update the weights of the model, a number of samples from the training dataset must be used to calculate the error: this hyperparameter is denominated batch size. The batch size is a hyperparameter that defines the number of samples to go through before updating the internal model parameters. The batch size has an effect on the resource requirements of the learning process. A large batch size (sometimes equal to the entire dataset) allows computational boosts, but at the expense of more memory for the training process. A small batch size is advantageous since it prevents the training process from stopping at local minima, but it is also considerably more prone to noise during the error calculations. The optimal size will depend on many factors, including the available memory capacity, as well as the size of training dataset.

Once the loss gradient has been estimated, the derivative of the error can be calculated and used to update the weights of the network through:

$$w_i^{new} = w_i^{old} - \gamma * \frac{\partial L}{\partial w_i} \tag{2.2}$$

where w_i^{new} and w_i^{old} are the current and previous weights respectively, $\frac{\partial L}{\partial w_i}$ is the gradient of the loss function (in this one-dimension case, it will be the first derivate of loss function with respect to only one weight) and γ is the learning rate. The learning rate controls at which pace the weights are updated. The optimal value for the learning rate depends on the problem under study. In general, if it is too large, the learning process is fast - since big steps are being made at each iteration, however, in this case, the minimum value is likely to be missed, making it almost impossible for the algorithm to converge. Conversely, if the learning rate is small, small advances will be made at each iteration, improving the likelihood of smoothly reaching a local minimum. Nevertheless, this can cause the learning process to be too slow.

Finally, the training process must be repeated several times until a good or good enough set of model parameters is discovered. The total number of iterations of the process is bounded by the number of complete passes through the data after which the training process is complete. This is referred to as the number of epochs. The optimal number of epochs can be reached using, for instance, a strategy of imposing a condition, e.g. epochs stop when the error is close to zero.

2.3.2.2 Datasets, Overfitting and Regularisation

The ultimate goal of a neural network is to have a model which performs well both on the data used to train and the new, previously unseen data on which the model will be used to make predictions. The ability to perform well on previously unobserved inputs is called generalisation. In order to accomplish such purpose, firstly, the dataset should be divided into three partitions - training set, validation set and test set. The training dataset is the sample of data used to fit the model (find the weights and biases, as previously described). The validation set is used to evaluate a given model and fine-tune the model hyperparameters, providing a proxy measure of model accuracy during the optimisation process. If the performance of the model on validation data is satisfactory, then its performance can be measured on the test dataset, which the model has never seen before this point. The results obtained with the test dataset are the gold standard used to evaluate the model. This strategy enables a fair evaluation of the model by measuring how well it generalises on new data unseen by the model.

According to not only the total number of samples, but also the actual developed model, different dataset split ratios might be employed. 60:20:20 or 75:15:10 are common ratios for training, validation and test, respectively. However, many times, instead of splitting the data into three different groups, they are only divided into two datasets - train and test. After this, the train dataset is split into X% to be the actual train set and the remaining (100-X)% is determined to be the validation set, where X is a fixed number (e.g. 80%). The model is then iteratively trained and validated on these different sets. This procedure is known as cross-validation (the basic form of cross-validation is k-fold cross-validation). The prime reason for the use of cross-validation, instead of the conventional validation, is related to insufficient data samples, making infeasible to partition these into separate training, validation and test sets. Additionally, the cross-validation procedure ensures that an unbiased model is produced.

The result of running the trained-validated model on the test set, will ideally be an accurate model which is able to generalise well on new data, i.e. provides a good fit to the data. A model with these characteristics is neither underfit nor overfit. An underfitting model is not able to adequately capture the underlying structure of the data and performs poorly in both the training and test sets. Effective procedures to avoid underfitting can easily be employed: for instance, increasing the capacity of the model by changing its structure adding more layers and/or more nodes to layers. Conversely, a model is overfitting if it finds a solution that is too tightly linked to the training examples. Instead of capturing the dominant trend of the data, it learns too many patterns and characteristics from the training data, which are not required for the output prediction.

Avoiding overfitting can be achieved through different regularisation methods, such as obtaining more data or using data-augmentation approaches to artificial generate more samples from the training data. L2 regularisation algorithm is, however, the most popular technique implemented to reduce overfitting. Specifically, a penalty term - called weight decay: (λ)- is added to the loss function on the training set to reduce the complexity of the learned model. This value controls the penalisation added to the model, based on the size of the weights allowing the model to build solutions based on multiple features, while limiting its flexibility and improving its generalisation capacity.

Additionally, dropout, which consists of ignoring (or dropping) a random fraction of units and their corresponding connections, has the power to reduce interdependent learning among

neurons, forcing the model to avoid relying excessively on particular set of features, thus being extensively used to avoid overfitting.

2.3.3 Evaluation Metrics

An essential element of deep learning algorithms is the selection of the right evaluation metrics. A suitable metric allows the comparison of the performance of different algorithms, while limiting the bias in the reported results. It is standard to provide multiple metrics, nevertheless, algorithms are generally evaluated according to, among others, the function they were aimed at.

Specifically for classification tasks, the confusion matrix - Figure 2.15 - is an intuitive and easy manner to report results, which displays all possible outcomes regarding the class prediction.

	True label positive	True label negative
Predicted label positive	# of true positives, TP	# of false positives, FP
Predicted label negative	# of false negatives, FN	# of true negatives, TN

Figure 2.15: Confusion matrix for a binary classifier.

The most commonly used metrics can be calculated based on the confusion matrix, such as accuracy (Equation 2.3), precision (Equation 2.4), recall or sensitivity (Equation 2.5) and specificity (Equation 2.6).

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{2.3}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.4}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.5}$$

$$Specificity = \frac{TN}{TN+FN} \tag{2.6}$$

Although confusion matrix can be applied both to the classical binary model and to the multi-class model, the application of the above referred metrics is restricted to binary classifiers. F1 Score - Equation 2.7 - is the weighted average of precision Equation 2.4 and recall Equation 2.5. Therefore, this score takes both false positives and false negatives into account. Intuitively, it is not as easy to understand as accuracy, but F1 score is usually more useful than accuracy, especially in presence of uneven class distribution.

$$F_1 score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.7}$$

Cohen's Kappa score is an evaluation statistic that takes into account how much agreement would be expected by chance and is commonly computed in the presence of unbalanced datasets.

Its values range from 0 - no agreement, other than what would be expected by chance - to 1 - complete agreement. This coefficient can also take negative values, which implies that there is no effective agreement between the set of values or the agreement is worse than random.

2.3.4 Convolutional Neural Networks (CNN)

CNN are a variant of neural networks that use convolution operations in, at least, one of the hidden layers. These networks are specifically designed to process large-scale images or sensory data in the form of multiple arrays [92]. The impressive performance of CNN and its great success in the field of computer vision and pattern recognition [93], such as for visual recognition [94], image retrieval [95] and scene annotation [96] has propelled its use and continuous development and optimisation. The general CNN architecture is represented in Figure 2.16: it is composed of different types of layers: Convolutional layers, Pooling layers and Fully Connected layers - these last especifically for classification tasks.

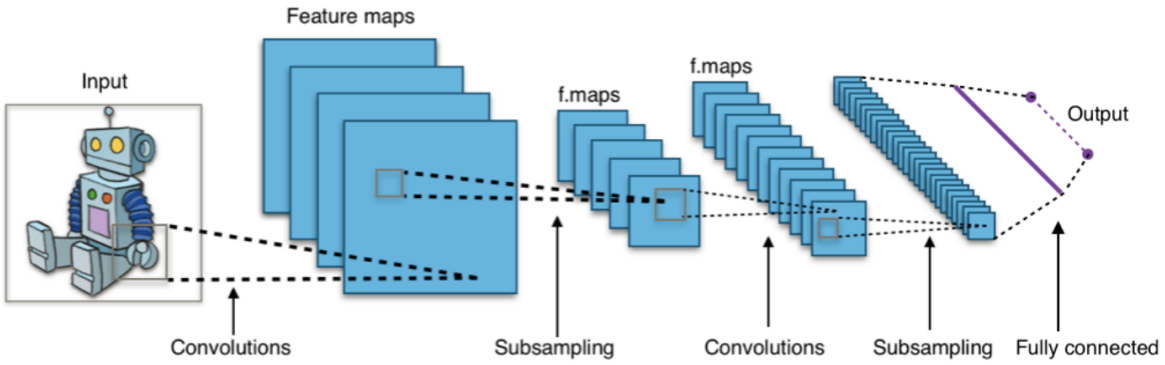


Figure 2.16: Basic CNN architecture composed of convolutional layers, followed by subsampling, e.g. pooling, and fully connected layer. The width of the model increases as the depth grows. Adapted from [97].

The real advantage of deep learning, especially for image recognition, comes from convolutional layers, since it is responsible for capturing features from the images, which is the essence of a CNN. The first layers usually obtain low-level features (like edges, lines and corners) while the others get high-level features (like structures, objects and shapes).

The process made by convolutional layers can be decomposed into two phases: (i) the convolution step, where a fixed-size window, named kernel, is convolved over the image. Due to the fact that these kernels are spatially smaller than the input image, a parameter named stride controls the sliding of the kernels through all the image, defining a region of interest. For instance, by defining a stride equal to one, kernels are moved one pixel at a time, while for a stride equal to two, kernels advance two pixels at a time, as they are being slid through the input. (ii) The processing step uses the pixels inside each window as input for the neurons, which finally perform the feature extraction from this specific region. Formally, in the latter step, each pixel is multiplied by its respective weight, generating the output of the neuron. Thus, only one output is generated concerning each region defined by the kernel. This iterative process results in a new image (or feature map), generally smaller than the original one, with extracted visual features. An illustration summarising this process is presented in Figure 2.17.

Many of these features extracted from the convolutional layers are very similar, since each kernel may have common pixels, generating redundant information. Therefore, convolutional layers are usually followed by pooling layers, which are aimed at simplifying or reducing the

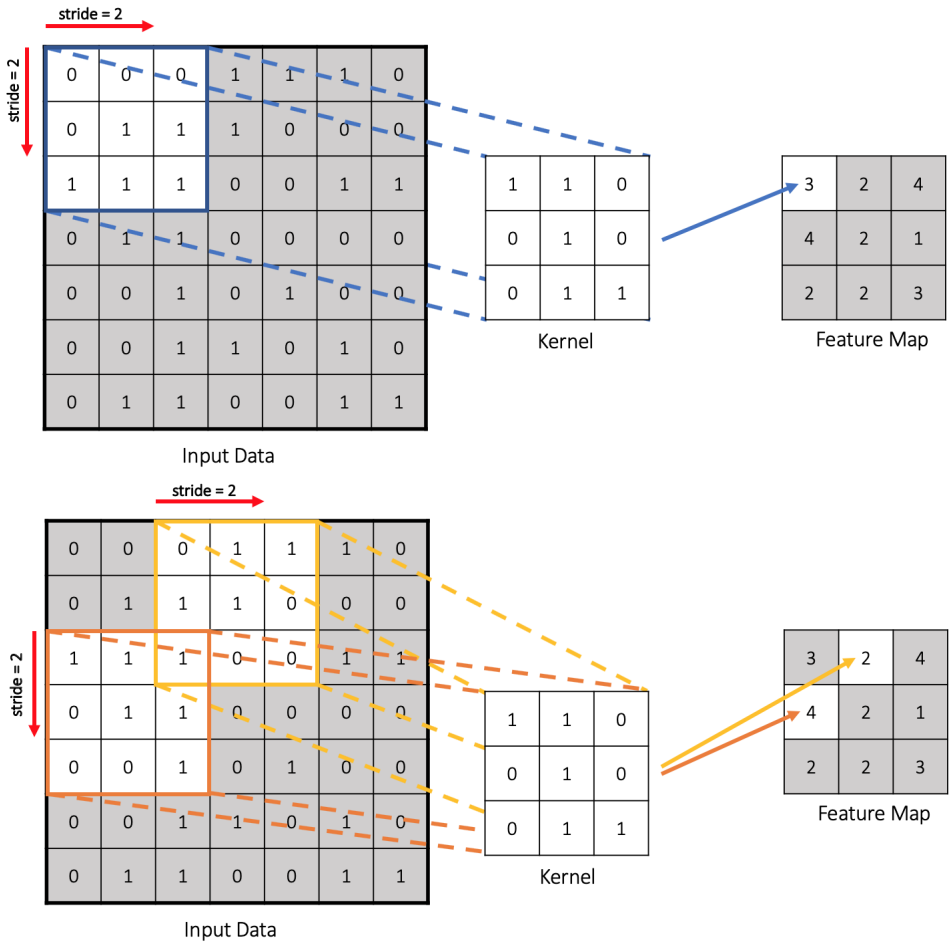


Figure 2.17: Convolution process with a kernel of 3× 3 and a stride of 2. The kernel slides through the input and at each position an element-wise multiplication and addition is performed, originating the feature map of size 3× 3.

spatial dimensions of the information derived from the feature maps, thus, preventing overfitting and improving its computational efficiency. Pooling layers are, hence, also known as downsampling layers. Specifically, in this type of layer a filter (and also a stride of the same size) is applied to the input feature map and the pooling function replaces the input with a summary statistic of the nearby feature parameters, creating a condensed feature map. The most commonly used pooling methods include the average pooling and the max pooling [98]. The former applies an average function to the input feature map, while the latter employs a maximum function (Figure 2.18). In all of these cases, pooling makes the network less sensitive to small translations of the input images, since it considers whole regions instead of separate input values, providing emphasis to the feature weight regardless of its location.

After several convolution and pooling layers, the CNN are, generally, finalised with a fully connected layer for classification tasks. This type of layer takes all neurons in the previous layer and connects them to every single neuron in its layer. Fully connected layers perform like a traditional neural network and contain about 90% of the parameters in a CNN. Its function is to take as input the output of the previous pooling layer, and produce an N dimensional vector, where N corresponds to the number of possible predicted classes within the model.

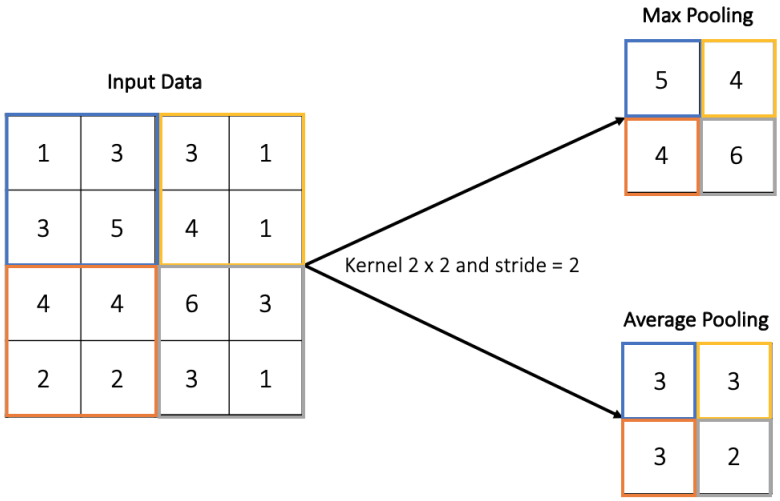


Figure 2.18: Schematic illustration of a 2× 2 pooling function. Each region of the input data is replaced by its maximum or average value, for max and average pooling, respectively.

3 MATERIALS AND METHODS

This chapter aims to describe the materials and methods used to ultimately develop the deep learning models used to predict PCa and BCR. Firstly, a description of the data, as well as its collection and the used processing techniques are presented. Then, the two different classification models are discussed. The first model is used to predict the following classifications: normal *versus* tumour *versus* high-grade PIN, while the second model concerns the prediction of PCa non-recurrence *versus* PCa recurrence.

3.1 Data Description

A retrospective pilot study was conducted randomly choosing 200 prostate cancer cases treated at the Champalimaud Clinical Center, between 2014 and 2017. The cohort comprises cases of men who, after being diagnosed with prostate cancer, were subjected to a radical surgery. Follow-up of these patients during the following two years after intervention, were also fully carried out at Champalimaud Clinical Center. Patients who received preoperative hormone or radiation therapy were excluded from this study. Postoperative treatments are only applied after the occurrence of relapse, thus if patients undergo additional therapies, it does not affect the research outcome. It is worth noting that no changes to the clinical standard procedures were introduced. This study was also approved by the Ethics Committee of Champalimaud Foundation. In order to obtain the above mentioned approval, several documents had to be written. These are included in the Appendix section.

For the first classification task, the 200 prostate cases were included. However, due to insufficient information regarding follow-up or other complications, 22 patients were excluded, resulting in a total of 178 cases for the second classification task.

3.2 Data Collection

As opposed to radiological images, which are generated within a computer, facilitating its posterior processing and analysis, histopathology slides images have to be digitised in order to become digitally available. Before the digitasion of the selected clinical cases, several stages, illustrated in Figure 3.1, had to be completed.

Firstly, H&E tissue sections corresponding to the whole prostate piece were retrieved from the pathology archive (around 40 glass slides per patient). Typically, the tumour was not restricted to one specific area, but instead distributed through different glass slides, since as demonstrated

by several authors [99–101], prostate cancer tends to be a multifocal disease. Thus, instead of selecting several images per patient - a laborious task - only the slides containing the index lesion were chosen for each of the patients. This procedure was applied because the index lesion - the largest tumor focus within the prostate - and its associated characteristics (namely, the grade and the presence or absence of extracapsular extension, lymph node invasion), generally drive the prognosis [101, 102]. Afterwards, the selected glass slides were digitised in a whole slide scanner (IntelliSite Ultra Fast Scanner, Philips [103]) and manually verified and downloaded from the Phillips Software Platform.

Figure 3.1: Necessary stages completed in order to transform the glass slide images into digital format.

For the first model, a total of 217 tissue sections were included, derived from 200 patients who underwent a radical prostatectomy, with a median of one tissue slide per patient and an interquartile range of 1 to 3. For the second model, 192 slides were included, derived from 178 patients, producing a median of one tissue slide per patient and an interquartile range of 1 to 3.

3.3 Whole Slide Image Processing

Image processing is of paramount importance, especially in the field of medical imaging. It allows enhancement and optimisation of the image characteristics such as sharpness and contrast. Hence, not only human viewers have an improved perception of the information contained within images, but also automated image techniques are provided with improved input images, offering both more sophisticated performance in distinctive tasks and the possibility of implementing methods which would be impracticable by alternative means.

3.3.1 Colour Normalisation

The appearance of the histological stains (e.g. H&E staining used in this study) often suffer from large variability [104]. While pathologists can effectively cope with staining variability, the performance of automated computational systems can be hampered by colour and intensity variations. Such variations in digital pathology images may be attributed to a number of factors, including specimen preparation and staining protocol inconsistencies (e.g. varying temperature of solutions), variations in fixation characteristics, inter-patient variation, and the scanner used to digitise the slides [104]. Despite the use of standardized staining protocols and automated staining machines which improve staining quality by yielding a more accurate and consistent staining, eliminating all the underlying sources of variation is not feasible [105]. Therefore, colour normalisation is an essential procedure to avoid discrepancies within the colours and intensities of the sample images.

Various colour normalisation approaches have been proposed in the literature [106–109]. However, these proposed algorithms have different shortcomings: they have been solely focusing on standardisation of patch images which contain only small regions of the whole image, disregarding the importance of colour relation between structures; they take fixed predefined

target colour images or use manual thresholds to establish the target model, making it harder to reproduce results.

These limitations were surpassed in [110], which presents a method, named Structure-Preserving Color Normalization (SPCN). SPCN works by replacing the colour basis of a source image - to normalise - with that of a pathologist-preferred target image (with the ideal colour and contrast), while still maintaining its original stain concentrations. The respective source code is available in [111] and was changed and adapted to be used in this dissertation.

3.3.2 Patching Approach: Patching the whole slide images

One typical whole histopathology section can be scanned to yield an image of a size larger than 100,000 × 100,000 pixels (usually up to 2 gigabytes with a JPEG compression scheme - default of the IntelliSite Ultra Fast Scanner software by Philips [103]), containing more than 1 million descriptive objects. Due to the inherent large-scale property of a histopathology image dataset, the feature extraction model needs to be both time and memory efficient, and the learning algorithm should be designed to be able to extract as much information as possible from these large images.

Although deep learning algorithms, specifically CNN, have the power to perform classification tasks [94], the extremely large size of a single histopathology image makes it unrealistic and impractical to construct a model taking these very large images as input. Based on this fact, and on all of the histopathology classification articles reviewed [87, 112–115], both classification frameworks presented in this study adopt a patch sampling technique to leverage CNN activation features of much smaller local patches, so that essential local details are preserved.

Hence, instead of providing the raw original whole slide image as input to the network, all the histopathologic images were patched. However, before this process, and only due to memory and time restrictions, the images had to be downsampled (approximately a 10× downsample). Afterwards, each histopathology image was divided into a set of non-overlapping smaller sub-images, i.e. square patches, with a size of 256× 256 pixels. The resulting sub-images, the patches or tiles, were in RGB, exactly like the original whole slide images. Of these generated RGB patches, some are composed solely of white background (for example in the corners and borders of the tiles, since the tissue tends to have an irregular format, in comparison to the field of view, which is rectangular). Thus, these did not contain any type of relevant information or features to be extracted and were discarded through an entropy-based elimination approach. In detail, this technique computes the entropy for each patch and eliminates it if the value is smaller than a specific threshold, which is empirically determined based on the intrinsic dataset characteristics.

3.3.3 Annotations

The generated patches - a total of 160,689 - were individually visualised and manually annotated by the author and reviewed by both the author and the Champalimaud Clinical Center expert pathologist (Prof. Dr. Antonio Beltran). Three different classes were distinguished: the first class was benign tissue which included both healthy glands and glands with low-grade PIN (class 0: negative); the second class was composed of tumour tissue (class 1: positive); and the third class comprises high-grade PIN instances (class 2: PIN). It is noteworthy, that when

high-grade PIN and malignant tissue were present within the same patch, the classification was always considered to be positive for tumour (i.e. class 1: positive).

As far as BCR prediction is concerned, the ground truth corresponded to the non-recurrent (class 0: negative) or recurrent (class 1: positive) condition which was obtained from the clinical information from each patient. A recurrence state was defined as a PSA value equal or greater than 0.2ng/dl, during the two years after the surgery date. From the 44,515 generated patches for this experiment, all the patches corresponding to a non-recurrent case were annotated as negative, while the patches corresponding to a recurrent case were annotated as positive. Table 3.1 summarises all of the information above.

Table 3.1: Number of cases, slides and generated patches for the two models developed in this study.

	Model 1: PCa Classification	**Model 2: PCa BCR Prediction**
Number of Cases	200	178
Number of Slides	217	192
Number of Patches	160,689	44,515

3.4 Deep Learning and CNN

Deep learning enables the extraction of multiple feature levels from data directly without explicit definition [116, 117]. CNN are the most successful deep learning models for processing multi-dimensional array data, such as colour images, being as a result the most common type of model used in digital pathology. Various CNN configurations have been developed AlexNet [118], VGGNet [119], GoogLeNet [120], Inception v3 [121]. Some of these architectures have repeatedly been shown to be successful for image recognition tasks and became standards in the field of digital histopathology [82, 112, 113, 115, 122, 123]. However, Inception v3 was the architecture chosen to be used in this study, since it has been shown to achieve great classification performance, especially for outcome prediction [67, 82, 83].

3.4.1 Inception V3 Implementation

In 2014, GoogLeNet, developed at Google, took first place in the 2014 ImageNet competition for classification and detection challenges [120]. Its main contribution was the development of an inception module, the basic unit of GoogLeNet. Within an inception module a series of convolutions at different scales are performed and subsequently concatenated into a single output vector. In order to save computation time, 1x1 convolutions are used to reduce the input channel depth. For each cell, a set of 1x1, 3x3 and 5x5 filters which learn to extract features at different scales from the input are used.

In order to further optimise the inception module, the same researchers developed a more efficient inception cell - Inception v3 [121]. Despite being beneficial in terms of their expressiveness and ability to extract features at a larger scale, convolutions with large spatial filters (such as 7x7) are computationally expensive, and were, thus, factorised into asymmetric convolutions: 1x7 and 7x1 convolutions. Additionally, 5x5 convolutions were replaced by two stacked 3x3 filters, providing a cheaper representation, while still maintaining the model performance. Auxiliary outputs were also added to the primary network, since they were found to benefit the

performance of the model, converging at a slightly better value than the same network architecture without an auxiliary part. The addition of auxiliary outputs also has a regularizing effect on the network, preventing overfitting.

As far as implementation is concerned, the Inception v3 architecture [124] was implemented in a platform specially developed and optimised for medical imaging applications, named NiftyNet [125]. NiftyNet is built using the TensorFlow [126] within the Python environment. It provides not only a high-level deep learning platform with components optimised for medical imaging applications (image segmentation, classification, regression, generation), but it also grants the tools to define and efficiently execute computational pipelines on hardware resources.

All of the deep learning models were executed in the Computational Cluster available at the Champalimaud Foundation. The machine specifications used to train, validate and test the developed algorithms are presented in Table 3.2.

Table 3.2: Hardware and Software specifications of the compute machine.

CPU	Intel (R) Xeon (R) CPU E5-2620v3 @ 2.40GHz
RAM	128GB DDR4 2133MHz
OS	CentOS release 7.7.1908 - kernel 3.10.0-1062.1.1.el7
GPU	NVIDIA TITAN X Pascal
CUDA	CUDA 10.1
Python	v3.6.8
Tensorflow	v1.10.0
NiftyNet	v0.5.0

3.4.2 Model I: PCa Classification

Given a patch as an input, the goal of this task was to classify the image into one of the three groups - negative for tumour (class 0), positive for tumour (class 1) and PIN (class 2). In order to do so, the data was randomly split into two datasets: training, validation and test, accounting for 90% and 10% of the total data respectively. It is worth noting that all tiles of the same patient were not separated, but were instead, associated as a whole to either the training/validation or test datasets to prevent contamination between the datasets and biases results. The resulting dataset is highly imbalanced, as shown in Table 3.3, specially class 2 which is around 45× less represented than class 0.

Table 3.3: Class representation for the dataset used to develop model I.

Classes	Number of Samples
Class 0	108,523
Class 1	49,735
Class 2	2,431

Firstly, an hyperparametresation of the network was performed using the training and validation data. In order to do so, some network parameters were modified in order to optimise the hyperparameters - specifically, learning rate, weight decay, number of epochs and batch size.

To find the appropriate parameters, diagnostic plots - which show loss over training epochs during training - were used to investigate how the different values of learning rate, batch size and number of epoch impact the rate of learning and learning dynamics of the model.

The learning rate (lr) is the most important hyperparameter to optimise, so that it balances the quality of the final solution with the training time required to achieve such outcome. A starting learning rate of 0.0001 was used, however, since this value made the network adjust too slowly, i.e. taking a long time to converge, a higher learning rate value (lr = 0.01) was used. This value was found to be too large, since the model appeared to converge too quickly, yet to a suboptimal solution. Therefore, a learning rate value of 0.001 was adopted because it represented a good trade-off value.

As far as weigh decay is concerned, its values range between 0.0 (no penalty) and 1.0 (full penalty). If the decay is high, the model will underestimate the weights and underfit the problem. Conversely, if the decay is too small, the model will be allowed to overfit the training data. Although a value of 0.01 tends to be high for most models, it was successfully adopted, allowing a well enough data fitted model.

The number of epochs is the number of times the entire training data is shown to the model. It plays an important role in how well the model fits the train data. Different values were tested, but 20,000 was the selected final value, given that the model converged and did not overfit. Values higher than 20,000 epochs deteriorate the model, since it, somehow, started to diverge and the accuracy and loss improvements were not significant.

Concerning the batch size, it is recommended to use batch sizes in powers of 2 (for better memory optimisation). Values of 16, 32, 64 and 128 were tested. Batch sizes of 16 and 32 originated a noisy loss function. Although a value of 128 was found to be better in terms of loss behavior, there were not enough memory resources to use this value. A batch size of 64 was selected, since a trade-off between time/memory resources and performance were required.

As far as training and validation are concerned, a k-fold cross-validation was used to reliably evaluate the model performance. The data was split into 5 partitions (k=5), instating 5 identical models and training each one of 4 (k-1) partitions, while evaluating on the remaining partitions. Figure 3.2 shows a schematic representation of a cross-validation procedure. The partition corresponding to the best model, i.e. the model which delivered an optimal performance was ultimately used on the test dataset. The performance of the models was assessed by using both visual tools - TensorBoard - and quantitative metrics - confusion matrix, accuracy and Cohen's Kappa statistic.

Model 1 was applied to both non-normalised and normalised (Subsections 4.2.1 and 4.2.2, respectively). These experiments provide interesting insights in terms of model behaviour, when visual improved images are provided.

3.4.3 Model II: PCa Biochemical Recurrence Prediction

Model 2 aims to classify each patch into one of the following groups: PCa non-recurrent (class 0) PCa recurrent (class 1). Due to the fact that relevant prognostic features are obtained from the characteristics of prostate adenocarcinoma [38], instead of using all the generated patches, only the tumour positive patches were considered for this second model. Majority voting among the predicted patch-wise labels of a given patient was used to categorize a case as non-recurrent or recurrent.

Data was divided into two datasets: training/validation (90% of the total dataset) and test (10% of the total dataset), while ensuring that tiles of each patient were contained only within one of the datasets. This procedure was applied due to the fact that an outcome per patient was intended to extract. and by having patches of the same patient split into different datasets this was not possible.

The class distribution for this dataset is presented below in Table 3.4.

Table 3.4: Class representation for the dataset used to develop model II.

Classes	Number of Cases	Number of Images
Class 0	132	26,806
Class 1	46	13,640

Before training, an hyperparametresation was performed. Different combinations of values for learning rate, weight decay, number of epochs and batch size were tested.

Concerning the learning rate, a value of 0.001 showed to deliver the best results. Smaller values were also tested, however they did not improve the performance of the model. The weight decay was chosen to be 0.0001, since it is within the range of values used for regularisation.

Given the complexity of the task in question, a higher value of epochs was chosen, specifically 35,000. Values of 15,000 and 60,000 were also tested. While the former value was too small, not allowing the model to converge, the latter yielded no significant improvements in terms of accuracy and loss, when compared to the value of 35,000.

Similarly to model I, the batch size was kept equal to 64, since it was the higher value that allowed training the model without failing due to lack of available memory. However, it is expected that higher batch size values could produce a noiseless model.

These above-mentioned parameter values were used for the training/validation and testing. Similarly to model 1, a 5 fold cross-validation approach was implemented Figure 3.2. From the 5 different partitions, the one with the best performance was selected to be used in the test dataset. In addition to the metrics used for model 1 - confusion matrix and Cohen's Kappa statistic - other metrics including accuracy, specificity, sensitivity, precision and F1 score were also used to assess the performance of this recurrence prediction model.

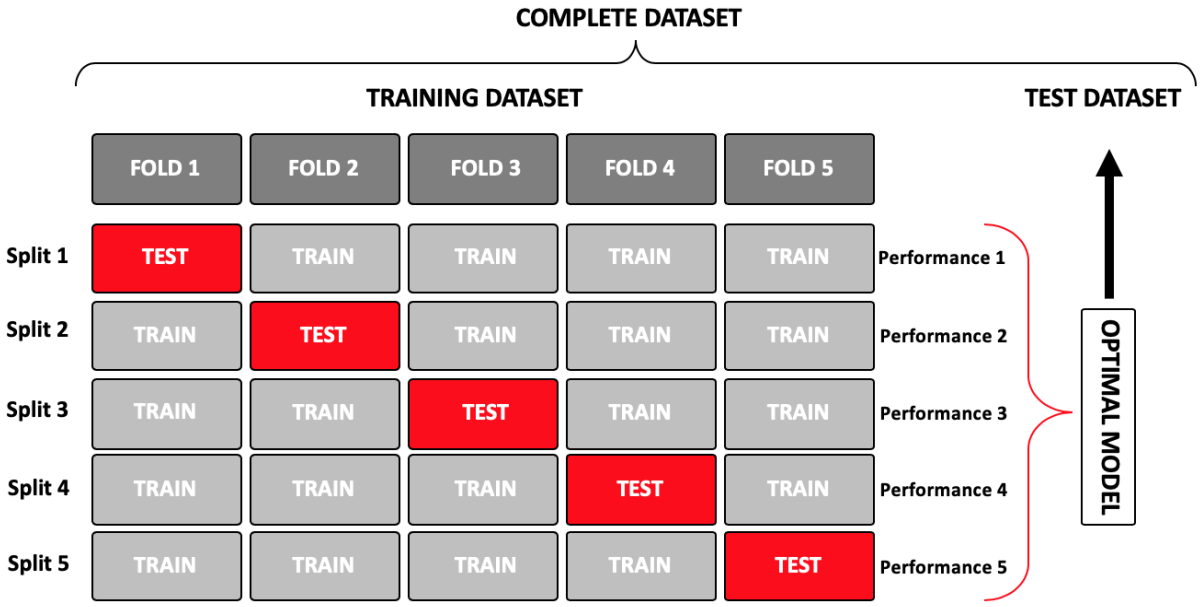


Figure 3.2: Overview of a k-fold cross-validation example, where k = 5. The entire dataset is partioned into 5 groups of samples, called folds.

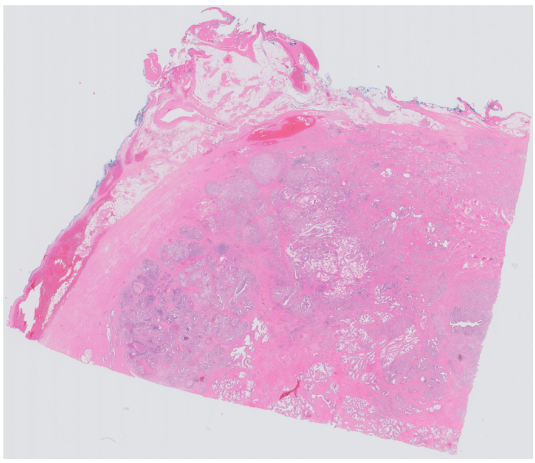
4 RESULTS

The major results are presented in this chapter. A brief discussion demonstrating the importance of image processing is, firstly, presented. Afterwards, the results from model I, both for non-normalised and normalised images are presented, as well as the results from model II. All the data analysis was performed using Python and all the developed code is available at [127].

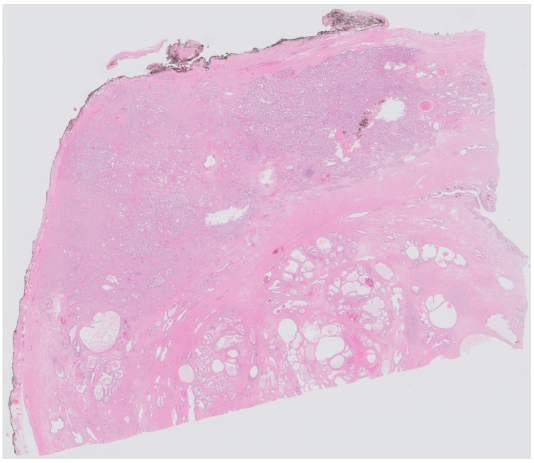
4.1 Importance of Image Processing

Variations in colour and intensity of H&E stained histological slides are granted features, when visualising this type of images. In order to improve consistency across the histologic images dataset, and ultimately improve classification effectiveness, a state of the art colour normalisation techniques was applied. An illustration of a whole slide image before and after normalisation is shown in Figure 4.1. Instead of only improving the image contrast, the original image structure i.e., the ratio between the different color concentrations, is preserved.

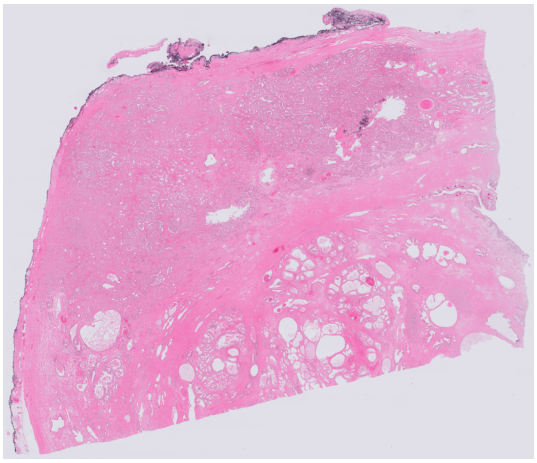
The patching approach phase is fundamental, as far as histological image analysis based on deep learning are concerned. A schematic illustration of the patch approach is presented in Figure 4.2. As can be seen, the whole slide image originates several patches that are only composed of white background. At the end of the entropy-based elimination process, which intended to discard these background patches, a mean of 741 patches per whole slide image was obtained, with a minimum of 143 and a maximum of 1304 patches per slide. Regarding the annotation phase, examples of the different classes considered are represented in Figure 4.3. All of these procedures culminated in the establishment of a database of prostate annotated histology images comprising 217 slides of 200 different cases.



(a) Target image.



(b) Original image.



(c) Normalised image.

Figure 4.1: Illustration of the performance of the colour normalisation technique. Given a selected target image (a), the original image (b) is transformed, yielding an image with improved contrast (c), i.e. the normalised image.

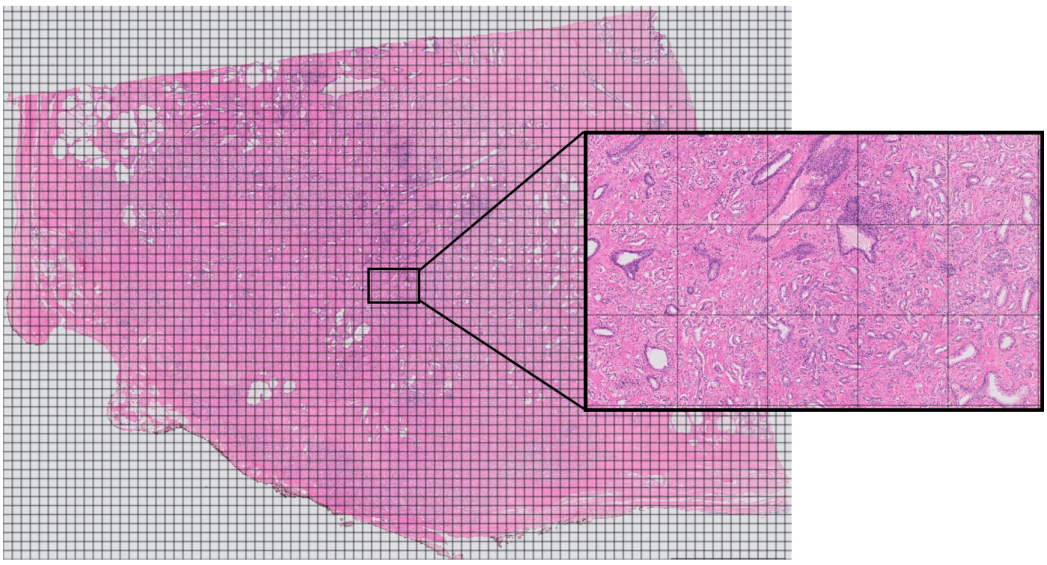


Figure 4.2: Schematic representation of the performed patching approach. The whole slide image generates multiple smaller images, which are used to feed the neural network.

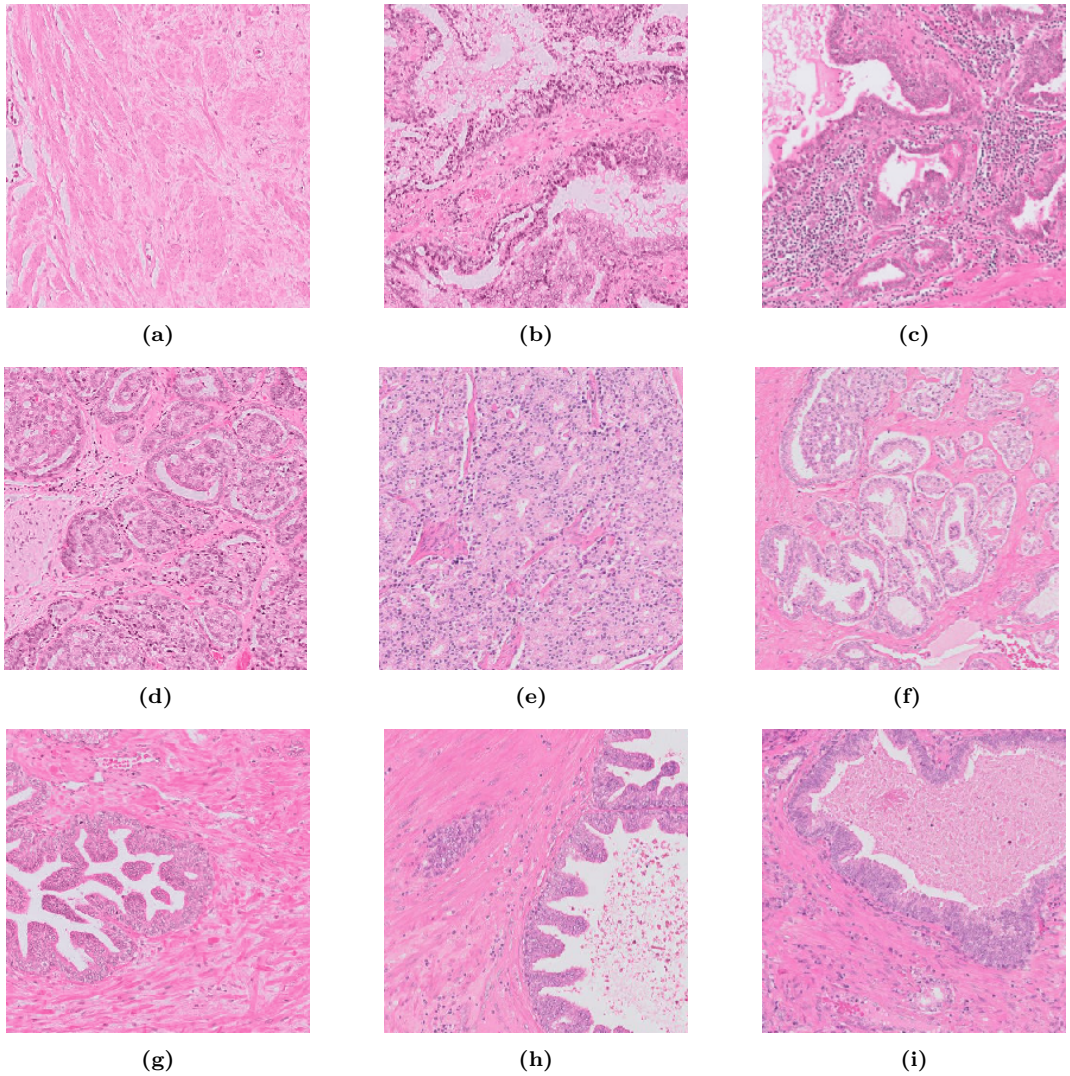


Figure 4.3: Different classes distinguished during the annotation process. (a), (b) and (c) are instances of benign tissue corresponding to class 0; (d), (e) and (f) are instances of malignant tumour corresponding to class 1; (g), (h) and (i) are instances of high-grade PIN corresponding to class 2.

4.2 Model I: PCa Classification

This multi-class task was performed for both non-normalised and normalised images. In order to accurately perform a comparison between the different approaches, the parameters of the network were kept constant and in the following setting:

Table 4.1: Learning parameters used to develop model I, both for non-normalised and normalised images.

Training Parameter	Value
Learning rate	0.001
Weight decay	0.01
Epochs	20,000
Batch size	64

4.2.1 Non-normalised images

Training and validation was performed following a 5-fold cross-validation in order to select the model yielding the best performance and use it for the test phase. Concerning quantitative evaluation parameters, since the data has an unbalanced nature (refer to Table 3.3), in addition to accuracy, Cohen's Kappa coefficient was also calculated. The values of the above mentioned metrics, resulting from the 5-fold cross-validation are shown in Table 4.2.

Table 4.2: Evaluation metrics calculated for 5 folds during the training/validation phase, concerning non-normalised images. The highlighted row corresponds to the fold yielding the best performance model.

	Accuracy	**Cohen's Kappa Coefficient**
Fold 1	0.94	0.85
Fold 2	0.93	0.84
Fold 3	0.92	0.81
Fold 4	0.93	0.81
Fold 5	**0.93**	**0.85**

The accuracy and loss, obtained from the best model - i.e. model produced by the fold 5 - are visualised in Figure 4.4 and Figure 4.5.

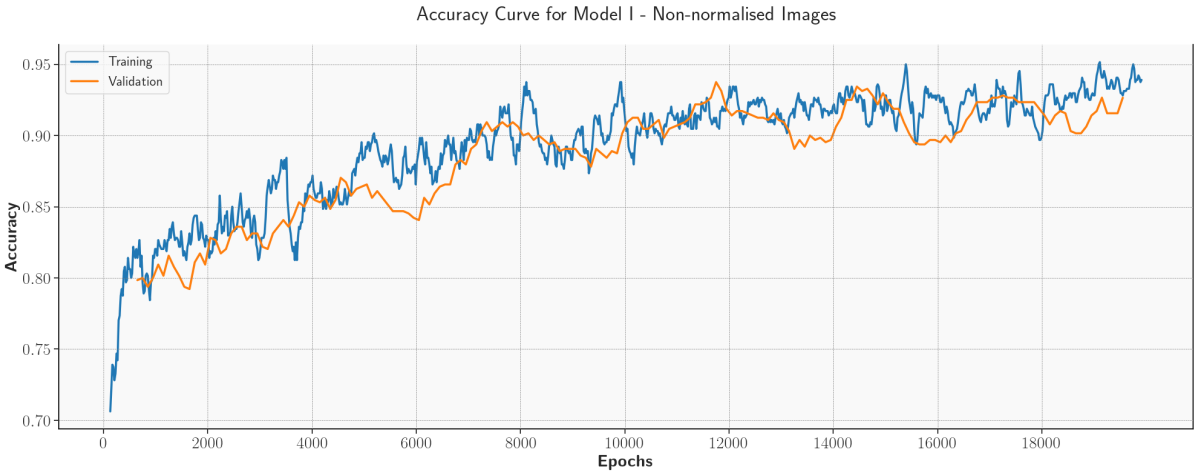


Figure 4.4: Accuracy progression for model I with non-normalised images. The blue line corresponds to training curve and the orange line to validation curve.

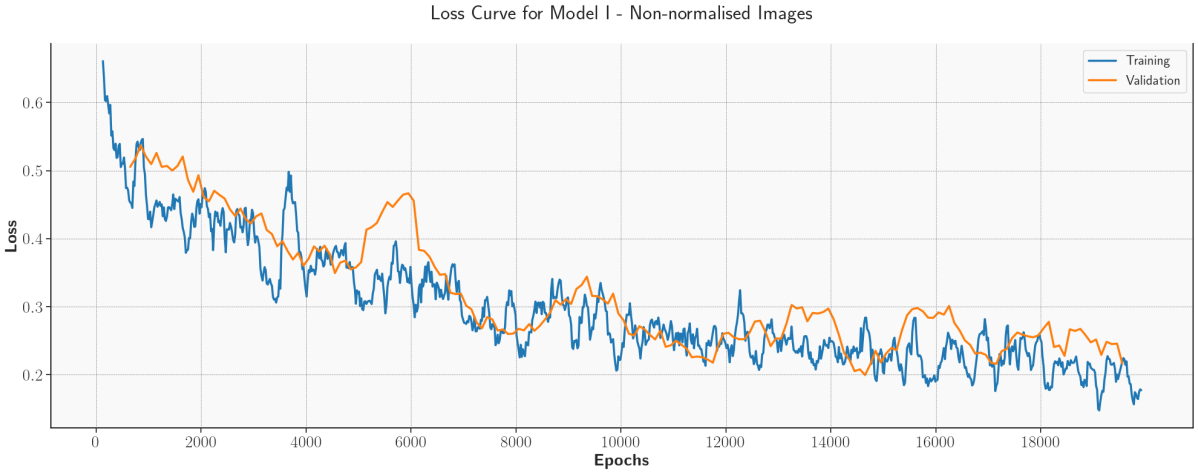


Figure 4.5: Loss progression for model I with non-normalised images. The blue line corresponds to training curve and the orange line to validation curve.

The final results, i.e the results for the test dataset, using the model obtained from fold 5, are presented below. The confusion matrix obtained for this three-class classification problem is illustrated in Figure 4.6. Instead of having the two entries found in the typical binary classification task, the confusion matrix has three entries, which correspond to the negative (class 0), positive (class 1) and PIN (class 2) conditions.

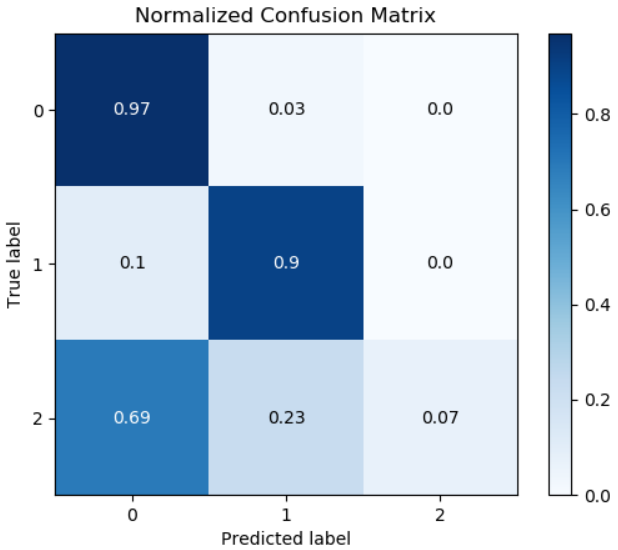


Figure 4.6: Normalised testset confusion matrix, corresponding to the results of model I for non-normalised images.

The accuracy value and the Cohen's Kappa coefficient were also determined and are presented in Table 4.3.

Table 4.3: Final values of accuracy and Cohen's Kappa coefficient yielded by the non-normalised images test dataset.

Evaluation Metric	Value
Accuracy	0.94
Cohen's Kappa Coefficient	0.85

In order to further quantify the quality of the model's prediction, regarding each of the classes at stake, and mainly because the classes are highly unbalanced, a binary analysis was performed. This is accomplished by treating the data as a collection of binary problems, one for each class. Some metrics which would only be defined for binary classification tasks can, hence, be determined.

Table 4.4: Different evaluation metrics calculated for the non-normalised images test dataset.

	Recall	Specificity	Precision	F1 Score
Class 0	0.97	0.88	0.94	0.96
Class 1	0.90	0.96	0.92	0.91
Class 2	0.07	0.99	0.59	0.13

4.2.2 Normalised images

Model I results, yielded by the normalised images are presented below. The values obtained both for accuracy and Cohen's Kappa coefficient, during the training/validation phase, with a 5-fold cross-validation strategy implemented are shown in Table 4.5.

Table 4.5: Evaluation metrics calculated for 5 folds during the training/validation phase, concerning normalised images. The highlighted row corresponds to the fold yielding the best performance model.

	Accuracy	Cohen's Kappa Coefficient
Fold 1	0.92	0.82
Fold 2	0.92	0.80
Fold 3	0.91	0.77
Fold 4	0.93	0.82
Fold 5	**0.94**	**0.85**

The graphics shown in Figure 4.7 and Figure 4.8 corresponded to the accuracy and loss respectively, obtained for the model which produced the best performance - in this case the model resulting from fold 5.

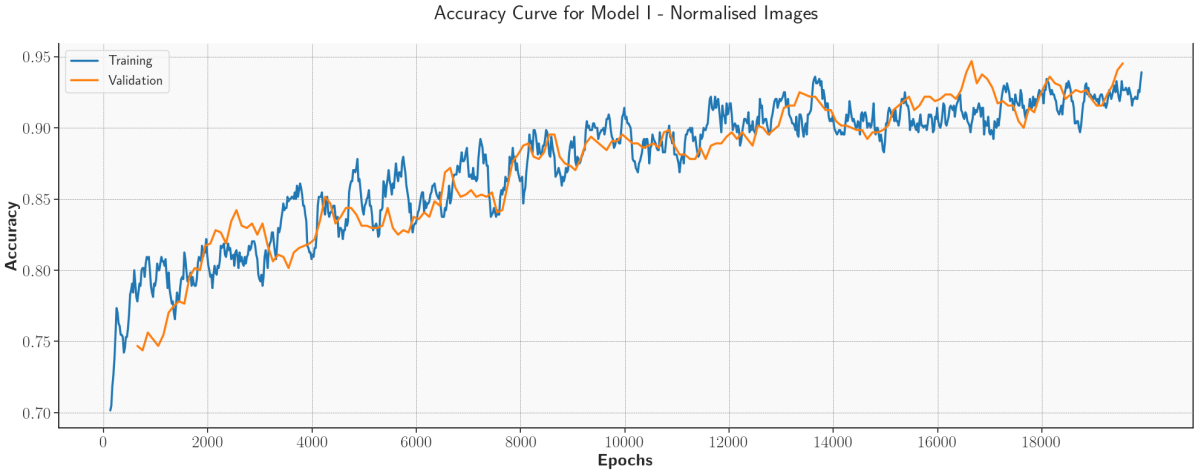


Figure 4.7: Accuracy progression for model I with normalised images. The blue line corresponds to training curve and the orange line to validation curve.

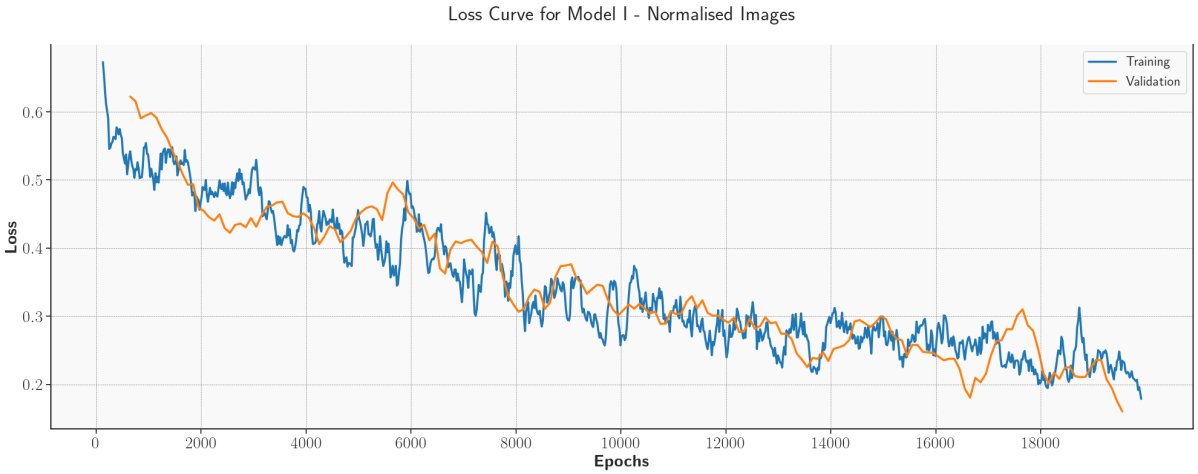


Figure 4.8: Loss progression for model I with normalised images. The blue line corresponds to training curve and the orange line to validation curve.

The results for the test dataset are presented below. The confusion matrix obtained for this three-class classification problem is illustrated in Figure 4.9.

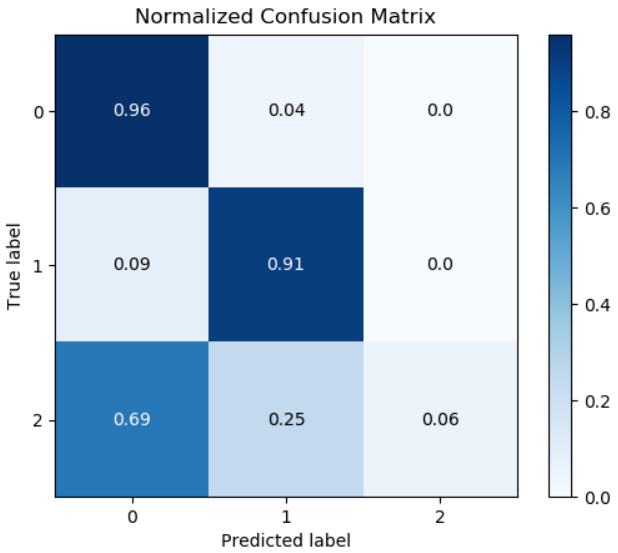


Figure 4.9: Normalised testset confusion matrix, corresponding to the results of model I for normalised images.

In Table 4.6, the final values of both accuracy and Cohen's Kappa coefficient are summarised.

Table 4.6: Accuracy and Cohen's Kappa Coefficient evaluation metrics for the testset, concerning normalised images.

Evaluation Metric	Value
Accuracy	0.93
Cohen's Kappa Coefficient	0.85

In addition to accuracy and Cohen's Kappa coefficient metrics, Table 4.7 presents the recall, specificity, precision and F1 score, which were calculated by treating the data as a collection of individual binary problems, one for each of the three classes considered.

Table 4.7: Different evaluation metrics calculated for the normalised images test dataset.

	Recall	Specificity	Precision	F1 Score
Class 0	0.96	0.90	0.93	0.94
Class 1	0.91	0.95	0.93	0.92
Class 2	0.06	0.99	0.56	0.10

4.3 Model II: PCa Biochemical Recurrence Prediction

Model II was aimed at differentiating between non-recurrent cases and recurrent cases, based on the tumour positive normalised patches (corresponding to the instances of class 1 in the previous model). The learning parameters used for this task are summarised in Table 4.8.

Table 4.8: Learning parameters used to develop model II.

Training Parameter	Value
Learning rate	0.001
Weight decay	0.0001
Epochs	35,000
Batch size	64

The accuracy and Cohen's Kappa coefficient metrics for the training/validation phase, with a 5-fold cross-validation implementation, are shown in Table 4.9. The highlighted row corresponds to the fold which produced the best performance model.

Table 4.9: Accuracy and Cohen's Kappa coefficient evaluation metrics calculated for the training/validation phase of model II.

	Accuracy	Cohen's Kappa Coefficient
Fold 1	**0.81**	**0.00**
Fold 2	0.80	0.00
Fold 3	0.56	0.00
Fold 4	0.65	0.00
Fold 5	0.49	0.00

The graphics corresponding to the accuracy and loss of the model produced by fold 1 are represented in Figure 4.10 and Figure 4.11, respectively.

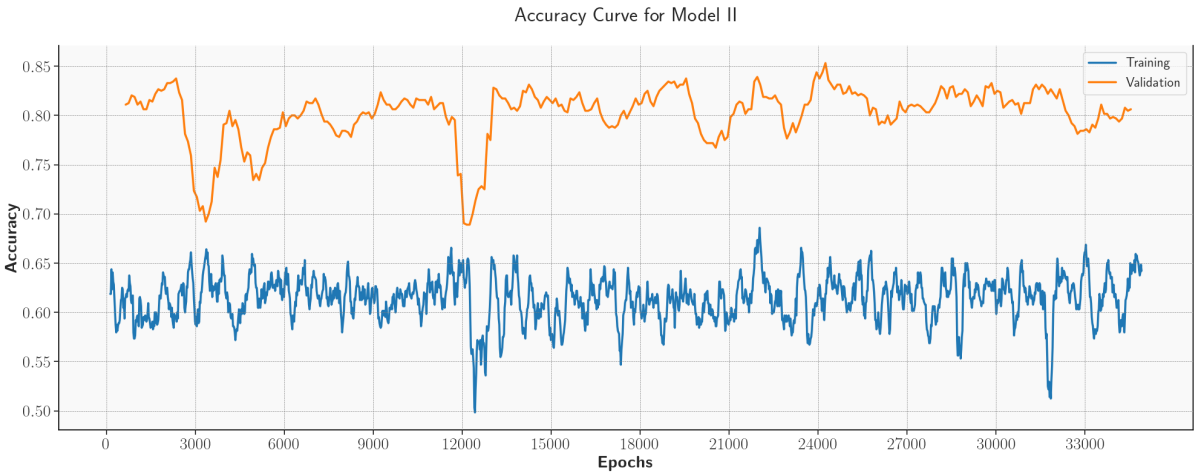


Figure 4.10: Accuracy progression for model II. The blue line corresponds to training curve and the orange line to validation curve.

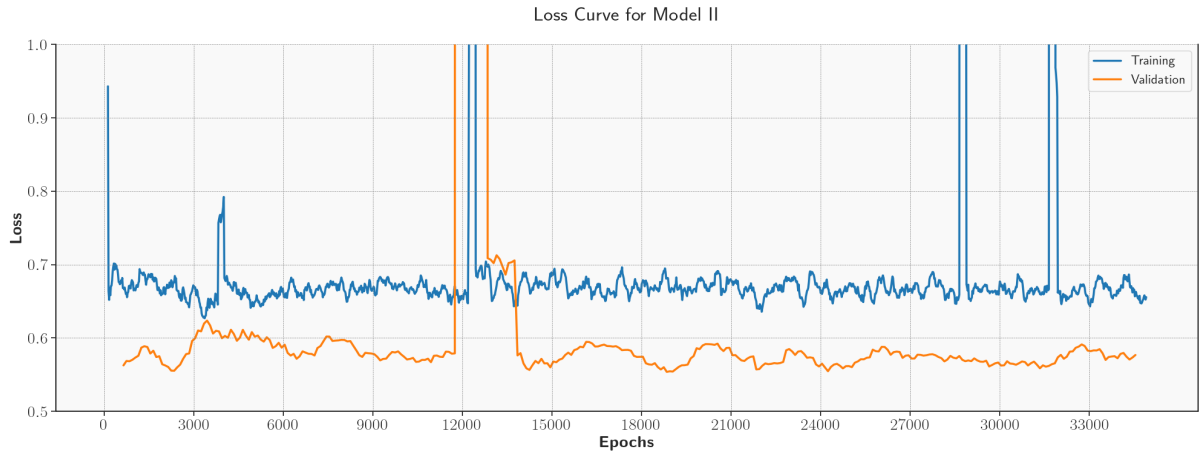


Figure 4.11: Loss progression for model II. The blue line corresponds to training curve and the orange line to validation curve.

Despite the fact the algorithm failed to learn and, thus, to correctly perform the proposed task, the whole deep learning pipeline was completed. Therefore, in Figure 4.12 the confusion matrix of the testset results is presented. Since this is a binary classification task, the confusion matrix follows the classical arrangement: it has two entries, corresponding to the negative class (class 0) and the positive class (class 1).

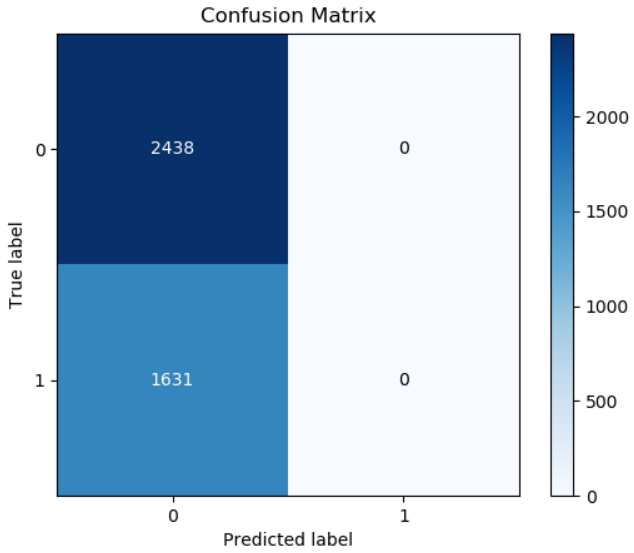


Figure 4.12: Testset confusion matrix, corresponding to the results of model II.

Majority voting was not applied to the results provided by the model, since all the patches were labeled as negative, which would result in classifying all the cases as negative, i.e. non-recurrent cases.

Based on the confusion, matrix some evaluation metrics were calculated - these are represented in Table 4.10.

Table 4.10: Different evaluation metrics calculated for the model II test dataset.

Evaluation Metric	Value
Accuracy	0.60
Cohen's Kappa Coefficient	0.00
Recall	0.00
Specificity	0.00
Precision	0.00
F1 Score	0.00

4.3.1 Resampling of the Original Dataset

Given the results obtained for the highly unbalanced BCR dataset, a resampling of the data was performed in order to understand if the justification for the model to fail was, somehow, related to the dataset class distribution. The characteristics of the new dataset, i.e. the resampled dataset are summarised in Table 4.11.

Table 4.11: Balanced dataset distribution.

	Class 0	Class 1	Total
Number of Cases	46	39	85
Number of Patches	13,326	13,326	26,648

Using the same learning parameters as for BCR original dataset, model II was trained and validated on the resampled dataset, producing the accuracy and loss curves represented in Figure 4.13 and Figure 4.14, respectively.

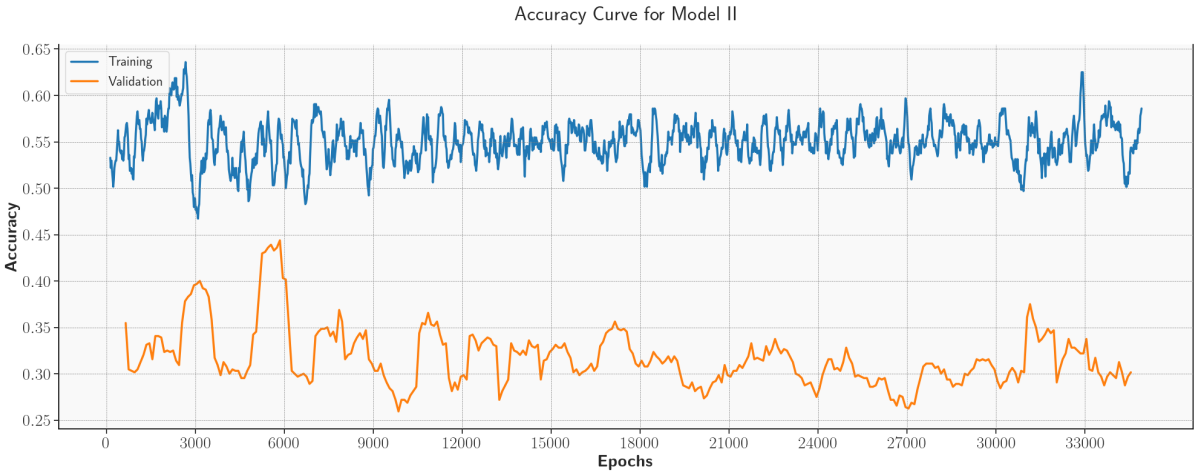


Figure 4.13: Accuracy progression for model II with the resampled dataset. The blue line corresponds to training curve and the orange line to validation curve.

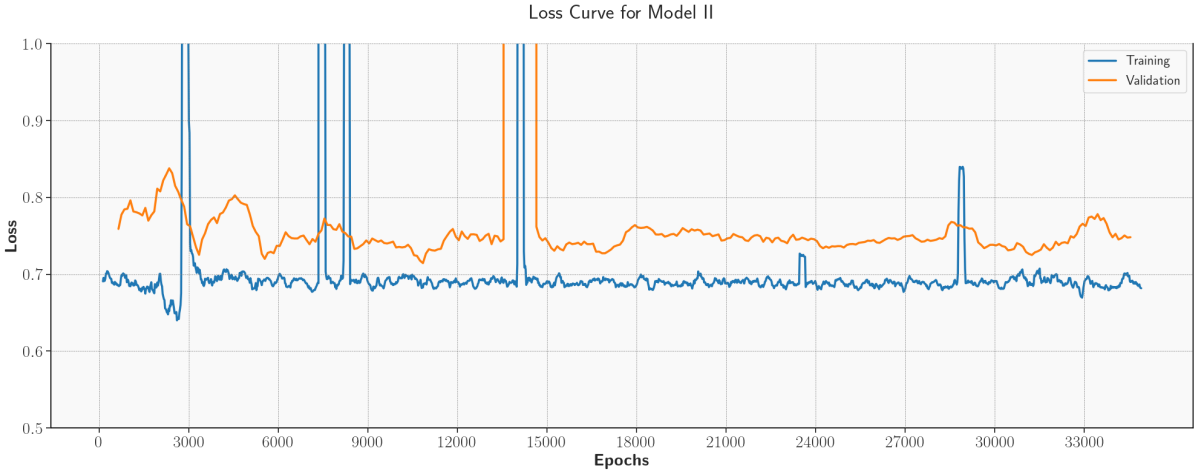


Figure 4.14: Loss progression for model II, with the resampled dataset. The blue line corresponds to training curve and the orange line to validation curve.

5 DISCUSSION

The application of image processing techniques to the H&E images allowed, as can be perceived in Figure 4.1, an improved contrast and also a decreased visual variability of the resulting tissue image dataset. Additionally, although pathologists are specifically trained to be able to cope with variations across slide images, the standardisation process allows an easier and straightforward annotation procedure, since the different structures can be easily differentiated. This is of extreme importance, specially when dealing with huge amounts of data, providing a faster labeling process, which within deep learning schemes takes a substantial amount of time. All of the collection and processing procedures - i.e. data curation - culminated in the establishment of an annotated database of prostate histological slides, with the corresponding clinical metadata. Apart from being used in this study to develop the deep learning models, the potential of this database is immense. A wide range of hypothesis can, now, be exploited and clinical questions regarding prostate cancer diagnosis, prognosis may be addressed.

In order to discern about the role of normalisation, i.e. if the improved visual quality of normalised images perceived by humans is translated into a better performance of a computer based algorithm, model I was developed both for non-normalised and normalised images. The results showed that little difference existed between the performance of the two models. Analysing, firstly, the training phase results, the values of the determined evaluation metrics for all the 5-fold cross-validation are very similar, both accuracy and Cohen's Kappa coefficient, specially for the best performance model (both yielded by fold 5). The value of the Cohen's Kappa coefficient is, however, smaller than accuracy, since this coefficient takes into account the fact that the different classes are highly unbalanced (refer to Table 3.3). Conversely, since the accuracy metric uses information of all class distributions of the dataset, is inherently sensitive to unbalanced distributions. Therefore, if class distribution varies - as from non-normalised to normalised test dataset - measures of the accuracy performance also change, even though the underlying fundamental performance of the classifier does not.

As far as accuracy and loss progression are concerned, it can be observed that, both curves follow the same pattern. The accuracy curves - both for non-normalised and normalised images, Figure 4.4 and Figure 4.7 respectively - show a convergence tendency for a value around 90%. The loss progress represented in Figure 4.5 and in Figure 4.8 respectively for non-normalised and normalised images, reveal a reduction of error after some iterations, with a convergence pattern. The combination of great accuracy with low loss accomplished by model I, corresponds to the best case scenario, when training deep learning models, specifically, CNN, since only small errors on few data are made.

The same results are also conveyed by the test dataset. While for the non-normalised images an accuracy value of 94% was obtained, a value of 93% was achieved by the model based on normalised images. The exact same value of Cohen's Kappa coefficient, 85%, was achieved by both models. Although it was expected that the normalisation procedure would improve the performance of the CNN, this did not happen. Hence, it can be concluded that for this specific problem, network and learning parameters, the normalisation procedure does not improve the inherently learning capacity of the developed model. This also reveals the robustness of the architecture Inception v3, for the applied learning parameters, which allows the model to cope with input, quantitatively and qualitatively different. However, it cannot be extrapolated that the same conclusions would be obtained if different networks architectures or even different learning parameters were implemented. For instance, if nuclei segmentation or tissue partitioning were the aimed tasks, higher resolution characteristics and features would be assessed, and it would be likely that the model would struggle to cope with variations in image appearance and staining, causing the algorithm to have a large number of errors for slightly differently stained images from the same structure.

Owing to the fact that both types of images provide similar models, as far as performance is concerned, only the detailed analysis of model I for normalised images will be discussed, since the same conclusions can be applied to the non-normalised images based model. Regarding the learning phase, it can be seen from Figure 4.7 and Figure 4.8 that the curves, although well behaved - i.e. showing a convergence pattern - present some noise. This is related to the maximum value of batch size allowed to be used (64), given the memory resources available. A higher value of batch size could deliver noiseless and more stable curves.

Figure 4.9, which summarises the results of the testset, is illustrative of the behaviour of the model for the different classes. Firstly, it can be noticed that the model fails to identify instances belonging to class 2 (only 6% are well classified). Instead it mainly identifies these images as instances belonging to class 0 (69%). This behaviour was, somehow, expected since this class was underrepresented as compared to classes 0 and 1. Consequently, the model is not able to learn specific PIN characteristics and features in order to accurately differentiate this category. Data augmentation, for class 2 could have been implemented, in order to increase the number of PIN instances, through different operations such as random rescaling, horizontal flips, padding, perturbations to brightness, contrast, and colour, as well as random cropping, which are commonly used to train large neural networks. However, the number of artificial generated images would be at least 20× higher than the *real* initial images and this type of process could highly influence and suppress or even change the intrinsic characteristics of the original class 2 images, thus, it was not implemented. Concerning classes 0 and 1, the model performs exceptionally well, achieving 96% and 91% of accurately classified instances, respectively. These values correspond to the recall, i.e. the effectiveness of the model when classifying positive labels (refer to Table 4.7). Specificity indicates how effectively a classifier identifies negative labels and its value is almost 100% for class 2, which is explained by the fact that only few instances are being correctly classified as positive and, therefore, all the remaining are classified as negative. The values of the same parameter for classes 0 and 1 are also high, which means that negative labels are correctly identified. Precision, a measure of exactness, provides insight into what proportion of positive predicted images are actually positive. For classes 0 and 1, the precision value is large (93%), meaning that there is a great chance of images to be whether negative or positive for tumour (the most represented classes of the model). Class 2 precision value is,

conversely, low - 56% - indicating that there are many false positives, considering the number of true positives. The above analysed metrics, similarly to accuracy, are sensitive to changes in data distributions. However, F1 score, a combination of precision and recall, measures the effectiveness of classification both for balanced and unbalanced class distributions. The values of this parameter to classes 0 and 1 are high - 94% and 92% - demonstrating the model's capacity to successfully discriminate tumour negative and positive instances, respectively. The same is not valid for class 2, for which the model achieves only a F1 score of 10%, indicating, once more, that the model is not capable of successfully learning the characteristics of the PIN class.

Given that the model was developed based on data, whose labeling was perform by humans (the author and the Champalimaud expert pathologist), and still achieved high results, means that the quality of the labeling of the data is also outstanding. Otherwise, the *garbage in, garbage out* thumb rule, commonly used in machine learning would dictate a poor performance of the model, since it would be difficult for the model to produce any kind of good results regardless of how good the labeling is.

In clinical practice it is of paramount importance to maximise the recall, since it essentially corresponds to the probability of detection of the studied characteristic. In this case, it is imperative not to miss positive patches, and in order to do so, class 1 recall should become 100%. Despite the fact that this value is only 91%, it still presents potential to be improved, which can be accomplished by increasing the number of instances belonging to this class, in order to balance the dataset and allow the algorithm to learn more characteristics of tumour positive class images. Regarding the PIN class (class 2), it was included in this classification problem due to the fact that specific types of cancer are, structurally, similar to high-grade PINs, being significant to differentiate between them. This type of distinction may gain extra interest and relevance if this classification model is used in images derived from biopsies, instead of from radical prostatectomies. In biopsies, differentiating between PIN and tumour might be an arduous exercise, since limited tissue slides are collected and the context of surrounding architecture and environment which are used to discern the possible classification of the structure is, most of the times, not accessible. It is expected that increasing the number of PIN instances, and therefore the images belonging to class 2, would lead to an improved result, enabling the application of the developed model to clinical practice.

Bearing in mind that (i) state of the art accuracy values for classification of histopathology images are above 90% and (ii) values for Cohen's Kappa coefficient above 80% are defined as almost perfect, it is possible to confidently conclude that model I performance is extremely good. An user interface to implement the developed model into clinical care has already been considered. The implementation of such methods provides improved efficiency and consistency, greatly benefiting current clinical practice.

The goal of model II was to differentiate between non-recurrent and recurrent cases, through the analysis of tumour positive patches. This type of prediction, as previously mentioned, cannot be performed by humans and different techniques mainly based on nomograms are used to predict the occurrence of relapse. Considering the training/validation results of the presented model, it is possible to conclude that the algorithm was not able to identify features that could translate the desired output. Figure 4.10 and Figure 4.11 clearly demonstrate that the model does not have the ability to successfully complete the referred task. The former figure, corresponding to the accuracy progress during training, reveals a curve that is kept almost constant through the whole learning process (in this particular case 35,000 epochs was chosen to be learning end point).

Although higher than 50% - the chance of achieving a correct prediction on a binary classification problem - the accuracy value corresponds approximately to the class 0 representation percentage (about 60% to 70%). Additionally, the progress of the loss curve reveals an excessively high loss value and absence of not only a decreasing pattern, but also a convergence point, results that corroborate the prediction incapacity of the model. Based solely on the behaviour of model II during the training phase, it can be stated that it fails to learn from the data. Consequently, the Cohen's Kappa coefficient holds a value of zero - meaning that there is no agreement between the predicted and true labels. After achieving these results, several attempts were made to change the hyperparemeters, but even after experimenting different hyperparameters combinations, the underlying result did not change.

Nevertheless, the pipeline was finished and the *"best"* performing model during training/validation was further used in the test dataset. Here, the term *"best"* refers simply to the model which had more instances belonging to class 0, producing, hence a misleading higher accuracy score. The confusion matrix - Figure 4.12 - produced by the testset results, as expected, demonstrates the model lack of success in predicting BCR. With this evaluation scheme, it is possible to note that the model classifies all the entries as belonging to class 0 and therefore, the more images with class 0, apparent better results would be delivered. The evaluation parameters summarised in Table 4.10 demonstrated, once more, what has been said about model II. All metrics hold a value of 0, except accuracy. This is derived from the fact that the accuracy metric in this case does not provide adequate information on a classifier's functionality: for instance, in a simple situation, if a given dataset includes 5% of minority class examples and 95% of majority examples, by classifying every example to be a majority class example, an accuracy of 95% will be produced. Taken at face value, 95% accuracy across the entire dataset appears superb; however, on the same token, this description fails to reflect the fact that 0% of minority examples are identified. The same circumstances lead the accuracy value of the developed model to achieve a 60% of accuracy, even though the model completely fails to deliver correct predictions regarding the input images.

Moreover, in order to understand if the unbalanced dataset nature of model II was responsible for the poor performance produced, a resampling of the dataset was performed. The resulting dataset was composed of, approximately, the same number of images for the two classes (refer to Table 4.11). From Figure 4.13 and Figure 4.14 it can be concluded that, the model, once more failed to learn from the data provided. Therefore, the class distribution of the dataset does not influence the results of the developed model for BCR prediction.

Concluding, although model II has failed, and its derived results are poor, several assumptions can be drawn. In first place, it is interesting to note that the same architecture - Inception v3 - and data, i.e. same images, processed all in the same way, were applied to the two addressed problems. While for the first problem, regarding PCa classification, a model delivering state of the art was accomplished, the model obtained to predict BCR was not successful. This is a characteristic of deep learning based approaches, many unknowns take part in producing the final output. An essential consideration, when using deep learning is that models need to be built with significant information, and one of the main justifications for models to fail is lack of data. In this particular case, model II was developed using about 40,000 images, which albeit being a quite high number, it is not close to the hundred of thousand of images employed for model I, not even from the millions of images that are used for ImageNet to accomplish a wide range of different classification tasks. Specifically in a medical environment, this is presented as

an enormous obstacle: the inability to obtain not only data, but mainly enough data.

Not having the right data and/or not all the necessary data to be able to answer the initially asked question may also have been one of the problems for the failure of the BCR prediction model. As already mentioned, the biochemical relapse prediction assignment is extremely complex. This is reflected on the few methods used to predict this condition, which are yet not as accurate and generalisable as intended. Moreover, these prediction techniques assume the integration of a variety of clinical characteristics that are recognised as having an important role on the BCR condition. Hence, it is likely that the proposed approach is too simple for the biochemical relapse prediction assignment. For instance, the introduction of clinical features, such as pre-operation PSA, Gleason score, SVI, extra capsular extension and margins status into the model might be advantageous, providing it with more information, which could, ultimately, be correlated to the extracted features of the histopathology images. Finally, most machine learning problems follow the *"No Free Lunch"* theorem which states that there is no one model that works best for every problem. The assumptions of a good model for one domain may not hold for another, so it is possible that the Inception v3 architecture may not be the best choice to the BCR prediction assignment.

6 CONCLUSIONS

PCa is one of the most common types of cancer, affecting millions of men, every year. The associated diagnosis and prognosis are highly depended on the histopathological analysis of the prostate specimen. The advent of digital pathology and the unfolding of diverse studies suggesting state of the art deep learning techniques applied to medical images, reveal the potential of combining these two techniques in order to achieve earlier detection, more accurate and consistent diagnosis and prognosis and consequently create an era of truly personalised medicine, which will ultimately improve patient's lives.

In this sense, the present dissertation intended to explore the potential of deep learning algorithms, specifically CNN, to harvest relevant data from histological images of the prostate. The ability of CNN to automatically learn appropriate representations of the data, using relatively little preprocessing and its independence from prior knowledge and human effort in feature design is a major advantage, when comparing these techniques to other image classification hand-engineered algorithms. CNN architectures are based on simple yet powerful concepts, and their mathematical implementation makes their design successful in a wide range of tasks.

The architecture of different networks were studied, yet Inception v3 was the choice to develop two different models. While the purpose of the first model was to classify images as negative or positive as far as the presence of prostate cancer was concerned; the second was intended to predict the occurrence of BCR.

However, in order to successfully develop deep learning models data curation is a major phase that must be completed. Therefore, an annotated database of histopathology images, composed of 200 prostate cancer cases who received surgical treatment, was established. Data annotation - performed both by the author and a pathologist – was a fundamental part of the deep learning projects, since it provided the setup for training the implemented architecture from scratch allowing it to accurately function.

The results obtained revealed state of the art accuracy for model I - 93% - and poor accuracy for model II - 60%. Given these results, one can conclude that model I could assist pathologists in the task of reviewing images obtained from radical prostatectomies and detect positive slides, for them to further review. In order to do so, the following is suggested: the development of an interface that can automatically perform all the data processing procedures and give an output per whole slide image - negative if all the patches are negative for tumour and positive if there is one, and only one, positive patch. Additionally, the performance shortcomings of model II would have to be surpassed. This could be accomplished by including prognostic clinical parameters or use all the patches both positive and negative, instead of only using the tumour positive patches.

Negative patches may, perhaps, have some kind of characteristic which can be correlated to the BCR prediction.

For future work, some limitations of the presented study should be addressed, in particular, the limited memory resources available, which dictated the downsampling of the whole slide images, decreasing their resolution and possibly the accuracy of the developed algorithms. Additionally, the developed models should be tested with external data, i.e. data from external institutions or data from repositories, such as TCGA. These types of experiments and tests will ultimately provide stronger evidence of the effectiveness and generalisation of the developed models.

Finally, bearing in mind all of the above mentioned, it is possible to concluded that the proposed goal for this project were fulfilled. Given the potential of the developed database and models, strategies to assure its continuous and increasing valorisation were also discussed. The application of these strategies will ensure the provision of a broader and more useful information for the definition of more consensual diagnosis and prognosis.

References

[1] S. J. Berry, D. S. Coffey, P. C. Walsh, and L. L. Ewing, "The development of human benign prostatic hyperplasia with age.," *The Journal of Urology*, vol. 132, no. 3, pp. 474–9, 1984.

[2] L. V. Kost and G. W. Evans, "Occurrence and significance of striated muscle within the prostate.," *The Journal of Urology*, vol. 92, pp. 703–4, 1964.

[3] P. Dasgupta, *ABC of Prostate Cancer (ABC Series Book 193)*. BMJ Books, 2011.

[4] J. D. Wilson, J. E. Griffin, M. Leshin, and F. W. George, "Role of gonadal hormones in development of the sexual phenotypes.," *Human genetics*, vol. 58, no. 1, pp. 78–84, 1981.

[5] O. S. Lowsley, "The development of the human prostate gland with reference to the development of other structures at the neck of the urinary bladder.," *American Journal of Anatomy*, vol. 13, no. 3, pp. 299–349, 1912.

[6] L. M. Franks, "Atrophy and hyperplasia in the prostate proper.," *The Journal of Pathology and Bacteriology*, vol. 68, no. 2, pp. 617–622, 1954.

[7] J. E. McNeal, "Anatomy of the prostate: an historical survey of divergent views.," *The Prostate*, vol. 1, no. 1, pp. 3–13, 1980.

[8] J. E. McNeal, "Normal histology of the prostate.," *The American Journal of Surgical Pathology*, vol. 12, no. 8, pp. 619–33, 1988.

[9] J. E. McNeal, "Normal and pathologic anatomy of prostate.," *Urology*, vol. 17, no. Suppl 3, pp. 11–6, 1981.

[10] A. G. Ayala, J. Y. Ro, R. Babaian, P. Troncoso, and D. J. Grignon, "The prostatic capsule: does it exist? Its importance in the staging and treatment of prostatic carcinoma.," *The American Journal of Surgical Pathology*, vol. 13, no. 1, pp. 21–7, 1989.

[11] A. Lopez-Beltran, L. Cheng, R. Montironi, and M. R. Raspollini, *Pathology of the Prostate*. Cambridge University Press, 2017.

[12] A. A. Villers, J. E. McNeal, E. A. Redwine, F. S. Freiha, and T. A. Stamey, "Pathogenesis and biological significance of seminal vesicle invasion in prostatic adenocarcinoma.," *The Journal of Urology*, vol. 143, no. 6, pp. 1183–7, 1990.

[13] M. Ohori, P. T. Scardino, S. L. Lapin, C. Seale-Hawkins, J. Link, and T. M. Wheeler, "The mechanisms and prognostic significance of seminal vesicle involvement by prostate cancer.," *The American Journal of Surgical Pathology*, vol. 17, no. 12, pp. 1252–61, 1993.

[14] P. C. Walsh and P. J. Donker, "Impotence following radical prostatectomy: insight into etiology and prevention.," *The Journal of Urology*, vol. 128, no. 3, pp. 492–7, 1982.

[15] J. A. Talcott, P. Rieker, K. J. Propert, J. A. Clark, P. W. Kantoff, K. I. Wishnow, K. R. Loughlin, and J. P. Richie, "Patient-Reported Impotence and Incontinence After Nerve-Sparing Radical Prostatectomy.," *JNCI Journal of the National Cancer Institute*, vol. 89, no. 15, pp. 1117–1123, 1997.

[16] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[17] The Global Cancer Observatory - International Agency for Research on Cancer, "Global cancer burden in 2018." `https://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf`. Accessed on 18/06/2019.

[18] Ministério da Saúde - Direção-Geral de Saúde, "Portugal, Doenças Oncológicas em números – 2014.," *Programa Nacional para as Doenças Oncológicas - 2014*, pp. 1–85, 2014.

REFERENCES

[19] Programa Nacional para as Doenças Oncológicas, "Programa nacional para as doenças oncológicas 2017.," *Direção-Geral da Saúde*, 2017.

[20] M. R. Cooperberg, J. M. Broering, P. W. Kantoff, and P. R. Carroll, "Contemporary trends in low risk prostate cancer: risk assessment and treatment.," *The Journal of Urology*, vol. 178, no. 3 Pt 2, pp. S14–19, 2007.

[21] G. Ploussard, J. I. Epstein, R. Montironi, P. R. Carroll, M. Wirth, M.-O. Grimm, A. S. Bjartell, F. Montorsi, S. J. Freedland, A. Erbersdobler, and T. H. van der Kwast, "The Contemporary Concept of Significant Versus Insignificant Prostate Cancer.," *European Urology*, vol. 60, no. 2, pp. 291–303, 2011.

[22] T. Wolters, M. J. Roobol, P. J. Van Leeuwen, R. C. N. Van Den Bergh, R. F. Hoedemaeker, G. J. L. H. Van Leenders, F. H. Schröder, and T. H. Van Der Kwast, "A Critical Analysis of the Tumor Volume Threshold for Clinically Insignificant Prostate Cancer Using a Data Set of a Randomized Screening Trial.," *The Journal of Urology*, 2011.

[23] W. Hamilton and D. Sharp, "Symptomatic diagnosis of prostate cancer in primary care: a structured review.," *The British Journal of General Practice : the Journal of the Royal College of General Practitioners*, vol. 54, no. 505, pp. 617–21, 2004.

[24] S.-M. Young, P. Bansal, E. T. Vella, A. Finelli, C. Levitt, and A. Loblaw, "Systematic review of clinical features of suspected prostate cancer in primary care.," *College of Family Physicians of Canada*, vol. 61, no. 1, pp. e26–35, 2015.

[25] M. J. Barry and L. H. Simmons, "Prevention of Prostate Cancer Morbidity and Mortality: Primary Prevention and Early Detection.," *The Medical Clinics of North America*, vol. 101, no. 4, pp. 787–806, 2017.

[26] J. V. Tricoli, M. Schoenfeldt, and B. A. Conley, "Detection of Prostate Cancer and Predicting Progression: Current and Future Diagnostic Markers.," tech. rep., 2004.

[27] N. Mottet, J. Bellmunt, M. Bolla, E. Briers, M. G. Cumberbatch, M. De Santis, N. Fossati, T. Gross, A. M. Henry, S. Joniau, T. B. Lam, M. D. Mason, V. B. Matveev, P. C. Moldovan, R. C. van den Bergh, T. Van den Broeck, H. G. van der Poel, T. H. van der Kwast, O. Rouvière, I. G. Schoots, T. Wiegel, and P. Cornford, "EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent.," *European Urology*, vol. 71, no. 4, pp. 618–629, 2017.

[28] M. Zhou and J. R. Srigley, "Benign mimickers and potential precursors of prostatic adenocarcinoma.," *Diagnostic Histopathology*, vol. 17, no. 10, pp. 434–446, 2011.

[29] J. I. Epstein, "Gleason score 2-4 adenocarcinoma of the prostate on needle biopsy: a diagnosis that should not be made.," *The American Journal of Surgical Pathology*, vol. 24, no. 4, pp. 477–8, 2000.

[30] T. H. van der Kwast, C. Lopes, C. Santonja, C.-G. Pihl, I. Neetens, P. Martikainen, S. Di Lollo, L. Bubendorf, R. F. Hoedemaeker, and m. o. T. P. C. O. T. E. R. S. O. S. F. P. C. Members of the pathology committee of the European Randomised Study of Screening for Prostate Cancer, "Guidelines for processing and reporting of prostatic needle biopsies.," *Journal of Clinical Pathology*, vol. 56, no. 5, pp. 336–40, 2003.

[31] A. Matoso and J. I. Epstein, "Grading of Prostate Cancer: Past, Present, and Future.," *Current Urology Reports*, vol. 17, no. 3, p. 25, 2016.

[32] J. R. Srigley, "Benign mimickers of prostatic adenocarcinoma.," *Modern Pathology*, vol. 17, no. 3, pp. 328–348, 2004.

[33] Y. Xu, Y. Wang, R. Zhou, H. Li, H. Cheng, Z. Wang, and J. Zhang, "The benign mimickers of prostatic acinar adenocarcinoma.," *Chinese Journal of Cancer Research*, vol. 28, no. 1, pp. 72–9, 2016.

[34] M. T. Rosenberg, E. S. Witt, M. Miner, and J. Barkin, "A practical primary care approach to lower urinary tract symptoms caused by benign prostatic hyperplasia (BPH-LUTS).," *The Canadian Journal of Urology*, vol. 21 Suppl 2, pp. 12–24, 2014.

[35] A. M. De Marzo, E. A. Platz, J. I. Epstein, T. Ali, A. Billis, T. Y. Chan, L. Cheng, M. Datta, L. Egevad, D. Ertoy-Baydar, X. Farree, S. W. Fine, K. A. Iczkowski, M. Ittmann, B. S. Knudsen, M. Loda, A. Lopez-Beltran, C. Magi-Galluzzi, G. Mikuz, R. Montironi, E. Pikarsky, G. Pizov, M. A. Rubin, H. Samaratunga, T. Sebo, I. A. Sesterhenn, R. B. Shah, S. Signoretti, J. Simko, G. Thomas, P. Troncoso, T. T. Tsuzuki, G. J. van Leenders, X. J. Yang, M. Zhou, W. D. Figg, A. Hoque, M. S. Lucia, A. Hoque, A. Hoque, and M. S. Lucia, "A Working Group Classification of Focal Prostate Atrophy Lesions.," *The American Journal of Surgical Pathology*, vol. 30, no. 10, pp. 1281–1291, 2006.

[36] M. K. Brawer, "Prostatic intraepithelial neoplasia: an overview.," *Reviews in Urology*, vol. 7 Suppl 3, no. Suppl 3, pp. S11–18, 2005.

[37] J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, P. A. Humphrey, and Grading Committee, "The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma.," *The American Journal of Surgical Pathology*, vol. 40, no. 2, p. 1, 2015.

[38] G. Sauter, S. Steurer, T. S. Clauditz, T. Krech, C. Wittmer, F. Lutz, M. Lennartz, T. Janssen, N. Hakimi, R. Simon, M. von Petersdorff-Campen, F. Jacobsen, K. von Loga, W. Wilczak, S. Minner, M. C. Tsourlakis, V. Chirico, A. Haese, H. Heinzer, B. Beyer, M. Graefen, U. Michl, G. Salomon, T. Steuber, L. H. Budäus, E. Hekeler, J. Malsy-Mink, S. Kutzera, C. Fraune, C. Göbel, H. Huland, and T. Schlomm, "Clinical Utility of Quantitative Gleason Grading in Prostate Biopsies and Prostatectomy Specimens.," *European Urology*, vol. 69, no. 4, pp. 592–598, 2016.

[39] D. G. Bostwick, M. B. Amin, P. Dundore, W. Marsh, and D. S. Schultz, "Architectural patterns of high-grade prostatic intraepithelial neoplasia.," *Human Pathology*, vol. 24, no. 3, pp. 298–310, 1993.

[40] W. B. Roberts and M. Han, "Clinical significance and treatment of biochemical recurrence after definitive therapy for localized prostate cancer.," *Surgical Oncology*, vol. 18, no. 3, pp. 268–274, 2009.

[41] S. Sridharan, V. Macias, K. Tangella, J. Melamed, E. Dube, M. X. Kong, A. Kajdacsy-Balla, and G. Popescu, "Prediction of prostate cancer recurrence using quantitative phase imaging: Validation on a general population.," *Scientific Reports*, vol. 6, pp. 1–10, 2016.

[42] J. L. Stanford, A. S. Hamilton, F. D. Gilliland, R. A. Stephenson, J. W. Eley, P. C. Albertsen, L. C. Harlan, and A. L. Potosky, "After Radical Prostatectomy for Clinically Localized Prostate Cancer The Prostate Cancer Outcomes Study.," vol. 1024, 2018.

[43] Y. Du, Q. Long, B. Guan, L. Mu, J. Tian, Y. Jiang, X. Bai, and D. Wu, "Robot-Assisted Radical Prostatectomy Is More Beneficial for Prostate Cancer Patients: A System Review and Meta-Analysis.," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 24, pp. 272–287, 2018.

[44] S. F. Shariat, M. W. Kattan, A. J. Vickers, P. I. Karakiewicz, and P. T. Scardino, "Critical review of prostate cancer predictive tools," *Future Oncology*, vol. 5, no. 10, pp. 1555–1584, 2009.

[45] J. F. Ward, M. L. Blute, J. Slezak, E. J. Bergstralh, and H. Zincke, "The long-term clinical impact of biochemical recurrence of prostate cancer 5 or more years after radical prostatectomy.," *Journal of Urology*, vol. 170, no. 5, pp. 1872–1876, 2003.

[46] G. W. Hull, F. Rabbani, F. Abbas, T. M. Wheeler, M. W. Kattan, and P. T. Scardino, "Cancer control with radical prostatectomy alone in 1,000 consecutive patients.," *The Journal of Urology*, vol. 167, no. 2 Pt 1, pp. 528–34, 2002.

[47] M. Moschini, V. Sharma, F. Zattoni, J. F. Quevedo, B. J. Davis, E. Kwon, and R. J. Karnes, "Natural History of Clinical Recurrence Patterns of Lymph Node–Positive Prostate Cancer After Radical Prostatectomy.," *European Urology*, vol. 69, no. 1, pp. 135–142, 2016.

[48] B. J. Miles, B. Giesler, and M. W. Kattan, "Recall and attitudes in patients with prostate cancer.," *Urology*, vol. 53, no. 1, pp. 169–174, 1999.

[49] A. S. Elstein, "Heuristics and biases: selected errors in clinical reasoning.," *Academic medicine: Journal of the Association of American Medical Colleges*, vol. 74, no. 7, pp. 791–4, 1999.

[50] P. L. Ross, C. Gerigk, M. Gonen, O. Yossepowitch, I. Cagiannos, P. C. Sogani, P. T. Scardino, and M. W. Kattan, "Comparisons of nomograms and urologists' predictions in prostate cancer.," *Seminars in Urologic Oncology*, vol. 20, no. 2, pp. 82–8, 2002.

[51] I. Vlaev and N. Chater, "Game relativity: How context influences strategic decision making.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 32, no. 1, pp. 131–149, 2006.

[52] P. L. Ross, P. T. Scardino, and M. W. Kattan, "A catalog of prostate cancer nomograms.," *The Journal of Urology*, vol. 165, no. 5, pp. 1562–8, 2001.

[53] M. W. Kattan, T. M. Wheeler, and P. T. Scardino, "Postoperative Nomogram for Disease Recurrence After Radical Prostatectomy for Prostate Cancer," *Journal of Clinical Oncology*, vol. 17, pp. 1499–1499, may 1999.

[54] A. J. Stephenson, P. T. Scardino, J. A. Eastham, F. J. Bianco, Z. A. Dotan, C. J. DiBlasio, A. Reuther, E. A. Klein, and M. W. Kattan, "Postoperative Nomogram Predicting the 10-Year Probability of Prostate Cancer Recurrence After Radical Prostatectomy.," *Journal of Clinical Oncology*, vol. 23, no. 28, pp. 7005–7012, 2005.

[55] M. R. Cooperberg, J. F. Hilton, and P. R. Carroll, "The CAPRA-S score.," *Cancer*, vol. 117, no. 22, pp. 5039–5046, 2011.

[56] L. Budäus, J. Schiffmann, M. Graefen, H. Huland, P. Tennstedt, A. Siegmann, D. Böhmer, V. Budach, D. Bartkowiak, and T. Wiegel, "Defining biochemical recurrence after radical prostatectomy and timing of early salvage radiotherapy.," *Strahlentherapie und Onkologie*, vol. 193, no. 9, pp. 692–699, 2017.

[57] K. L. Greene, M. V. Meng, E. P. Elkin, M. R. Cooperberg, D. J. Pasta, M. W. Kattan, K. Wallace, and P. R. Carroll, "Validation of the Kattan preoperative nomogram for prostate cancer recurrence using a community based cohort: results from cancer of the prostate strategic urological research endeavor (capsure).," *The Journal of Urology*, vol. 171, no. 6 Pt 1, pp. 2255–9, 2004.

[58] S. F. Shariat, P. I. Karakiewicz, C. G. Roehrborn, and M. W. Kattan, "An updated catalog of prostate cancer predictive tools.," *Cancer*, vol. 113, no. 11, pp. 3075–3099, 2008.

[59] G. Lughezzani, A. Briganti, P. I. Karakiewicz, M. W. Kattan, F. Montorsi, S. F. Shariat, and A. J. Vickers, "Predictive and Prognostic Models in Radical Prostatectomy Candidates: A Critical Analysis of the Literature.," *European Urology*, vol. 58, pp. 687–700, 2010.

[60] G. Ploussard, "Robotic surgery in urology: Facts and reality. What are the real advantages of robotic approaches for prostate cancer patients?.," *Current Opinion in Urology*, vol. 28, no. 2, pp. 153–158, 2018.

[61] C. Chan, A. W. Chiu, M. Chen, J. M. Hsu, S. Yang, and W. R. Lin, "A comparative study of laparoscopic and robotic assisted radical prostatectomy performed by a single surgeon.," *Urological Science*, vol. 28, no. 2, pp. 71–74, 2017.

[62] A. Gallina, P. I. Karakiewicz, G. C. Hutterer, F. K.-H. Chun, A. Briganti, J. Walz, E. Antebi, S. F. Shariat, N. Suardi, M. Graefen, A. Erbersdobler, A. Salonia, P. Rigatti, H. Huland, and F. Montorsi, "Obesity does not predispose to more aggressive prostate cancer either at biopsy or radical prostatectomy in European men.," *International Journal of Cancer*, vol. 121, no. 4, pp. 791–795, 2007.

[63] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, "Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images.," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[64] H. Mohan, *Textbook of Pathology/ Pathology Quick Review and MCQs*. Jaypee Brothers Medical Pub, 2010.

[65] J. K. C. Chan, "The Wonderful Colors of the Hematoxylin–Eosin Stain in Diagnostic Surgical Pathology.," *International Journal of Surgical Pathology*, vol. 22, no. 1, pp. 12–32, 2014.

[66] L. M. L. PhD, *Lippincott's Pocket Histology (Lippincott's Pocket Series)*. LWW, 2013.

[67] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep Learning for Identifying Metastatic Breast Cancer.," *ArXiv*, pp. 1–6, 2016.

[68] J. Ho, A. V. Parwani, D. M. Jukic, Y. Yagi, L. Anthony, and J. R. Gilbertson, "Use of whole slide imaging in surgical pathology quality assurance: design and pilot validation studies.," *Human Pathology*, vol. 37, no. 3, pp. 322–31, 2006.

[69] R. S. Weinstein, A. R. Graham, L. C. Richter, G. P. Barker, E. A. Krupinski, A. M. Lopez, K. A. Erps, A. K. Bhattacharyya, Y. Yagi, and J. R. Gilbertson, "Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future.," *Human Pathology*, vol. 40, no. 8, pp. 1057–1069, 2009.

[70] G. Bueno, M. Milagro Fernández-Carrobles, O. Deniz, and M. García-Rojo, "New Trends of Emerging Technologies in Digital Pathology.," *Pathobiology*, 2016.

[71] J. Griffin and D. Treanor, "Digital pathology in clinical use: where are we now and what is holding us back?.," *Histopathology*, vol. 70, no. 1, pp. 134–145, 2017.

[72] A. V. Parwani and L. Pantanowitz, "Symposium-New Frontiers in Digital Pathology.," *Journal of Pathology Informatics*, vol. 1, p. 15, 2010.

[73] Y. Sucaet and W. Waelput, *Digital Pathology (SpringerBriefs in Computer Science)*. Springer, 2014.

[74] G. O'Dowd, *Wheater's Pathology: A Text, Atlas and Review of Histopathology (Wheater's Histology and Pathology)*. Churchill Livingstone, 2019.

[75] Office of the Commissioner, "FDA allows marketing of first whole slide imaging system for digital pathology." `https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-whole-slide-imaging-system-digital-pathology`, April 12th, 2017. Accessed on 11/08/2019.

[76] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital Imaging in Pathology: Whole-Slide Imaging and Beyond," *Annual Review of Pathology: Mechanisms of Disease*, vol. 8, no. 1, pp. 331–359, 2013.

[77] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological Image Analysis: A Review.," *IEEE Reviews in Biomedical Engineering*, 2009.

[78] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities.," *Medical Image Analysis*, vol. 33, no. July, pp. 170–175, 2016.

[79] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model," *Scientific Reports*, vol. 7, no. 1, pp. 1–10, 2017.

[80] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 1455–1462, jul 2016.

[81] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt, "Toward a Shared Vision for Cancer Genomic Data.," *New England Journal of Medicine*, vol. 375, no. 12, pp. 1109–1112, 2016.

[82] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, A. Tsirigos, and P. Santiago Ocampo, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning.," *Nature Medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[83] D. Bychkov, N. Linder, R. Turkki, S. Nordling, P. E. Kovanen, C. Verrill, M. Walliander, M. Lundin, C. Haglund, and J. Lundin, "Deep learning based tissue analysis predicts outcome in colorectal cancer.," *Scientific Reports*, vol. 8, no. 1, pp. 1–11, 2018.

[84] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff, and M. Claassen, "Automated Gleason grading of prostate cancer tissue microarrays via deep learning.," *Scientific Reports*, vol. 8, no. 1, p. 12054, 2018.

[85] G. Nir, S. Hor, D. Karimi, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R. S. Wilson, K. A. Iczkowski, M. S. Lucia, P. C. Black, P. Abolmaesumi, S. L. Goldenberg, and S. E. Salcudean, "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts.," *Medical Image Analysis*, vol. 50, pp. 167–180, 2018.

[86] B. Acs and D. L. Rimm, "Not just digital pathology, intelligent digital pathology.," *JAMA Oncology*, vol. 4, no. 3, pp. 403–404, 2018.

[87] Y. Xu, Z. Jia, L.-B. Wang, Y. Ai, F. Zhang, M. Lai, and E. I.-C. Chang, "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features.," *BMC Bioinformatics*, vol. 18, p. 281, 2017.

[88] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Velázquez Vega, D. J. Brat, and L. A. D. Cooper, "Predicting cancer outcomes from histology and genomics using convolutional networks.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 13, pp. E2970–E2979, 2018.

[89] M. D. DiFranco, G. O'Hurley, E. W. Kay, R. W. G. Watson, and P. Cunningham, "Ensemble based system for whole-slide prostate cancer probability mapping using color texture features," *Computerized Medical Imaging and Graphics*, vol. 35, no. 7-8, pp. 629–645, 2011.

[90] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification.," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[91] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[92] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning.," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[93] J. Schmidhuber, "Deep learning in neural networks: An overview.," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," tech. rep., 2012.

[95] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval.," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 27–35, IEEE, 2015.

[96] E. Othman, Y. Bazi, N. Alajlan, H. Alhichri, and F. Melgani, "Using convolutional features and a sparse autoencoder for land-use scene classification.," *International Journal of Remote Sensing*, vol. 37, no. 10, pp. 2149–2167, 2016.

[97] B. B. Traore, B. Kamsu-Foguem, and F. Tangara, "Deep convolution neural network for image recognition.," *Ecological Informatics*, vol. 48, pp. 257–268, 2018.

[98] D. Scherer, A. Müller, and S. Behnke, "Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition," tech. rep.

[99] J. E. McNeal, E. A. Redwine, F. S. Freiha, and T. A. Stamey, "Zonal Distribution of Prostatic Adenocarcinoma.," *The American Journal of Surgical Pathology*, vol. 12, no. 12, pp. 897–906, 1988.

[100] A. Villers, J. E. McNeal, F. S. Freiha, and T. A. Stamey, "Multiple cancers in the prostate. Morphologic features of clinically recognized versus incidental tumors.," *Cancer*, vol. 70, no. 9, pp. 2313–8, 1992.

[101] M. Andreoiu and L. Cheng, "Multifocal prostate cancer: biologic, prognostic, and therapeutic implications.," *Human Pathology*, vol. 41, no. 6, pp. 781–793, 2010.

[102] H. U. Ahmed, "The index lesion and the origin of prostate cancer," *New England Journal of Medicine*, vol. 361, no. 17, p. 1704, 2009.

[103] Philips, "Intellisite ultra fast scanner, digital pathology slide scanner." `https://www.philips.co.uk/healthcare/product/HCNOCTN442/intellisite-ultra-fast-scanner`. Accessed on 22/07/2019.

[104] J. D. Bancroft, *Theory and Practice of Histological Techniques*. Churchill Livingstone, 2007.

[105] M. Niethammer, D. Borland, J. S. Marron, J. Woosley, and N. E. Thomas, "Appearance Normalization of Histology Slides," pp. 58–66, Springer, Berlin, Heidelberg, 2010.

[106] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images.," *IEEE Computer Graphics and Applications*, vol. 21, no. 4, pp. 34–41, 2001.

[107] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological Image Analysis: A Review.," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[108] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis.," *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, pp. 1107–1110, 2009.

[109] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution.," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1729–1738, 2014.

[110] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images.," *IEEE Transactions on Medical Imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[111] A. Vahadane, "Code release, color normalization." `https://github.com/abhishekvahadane/CodeRelease_ColorNormalization`, March 12th, 2016. Accessed on 10/07/2019.

[112] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases.," *Journal of Pathology Informatics*, vol. 7, no. 1, p. 29, 2016.

[113] N. Kumar, R. Verma, A. Arora, A. Kumar, S. Gupta, A. Sethi, and P. H. Gann, "Convolutional neural networks for prostate cancer recurrence prediction.," p. 101400H, 2017.

[114] S. Robertson, H. Azizpour, K. Smith, and J. Hartman, "Digital image analysis in breast pathology—from image processing techniques to artificial intelligence.," *Translational Research*, vol. 194, pp. 19–35, 2018.

[115] M. Lucas, I. Jansen, C. D. Savci-Heijink, S. L. Meijer, O. J. de Boer, T. G. van Leeuwen, D. M. de Bruin, and H. A. Marquering, "Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies.," *Virchows Archiv*, vol. 475, no. 1, pp. 77–83, 2019.

[116] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning.," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[117] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges.," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[118] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks,"

[119] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[120] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions.," 2014.

[121] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2015.

[122] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using Convolutional Neural Networks.," *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 2560–2567, 2016.

[123] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, L. Jansen, C. C. Reyes-Aldasoro, I. Zörnig, D. Jäger, H. Brenner, J. Chang-Claude, M. Hoffmeister, and N. Halama, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study.," *PLOS Medicine*, vol. 16, no. 1, p. e1002730, 2019.

[124] A. Vahadane, "Inception model." `https://github.com/tensorflow/models/blob/master/research/inception/inception/slim/inception_model.py`. Accessed on 17/05/2019.

[125] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "NiftyNet: a deep-learning platform for medical imaging," *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 113–122, 2018.

[126] "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[127] C. Seabra, "Github: Developed code for the dissertation project." `https://github.com/carolinaseabra/ProBCR-Project`, 2019.

Appendix

A.1 Ethics Committee Submission Documents

Study Summary

Title: Prostate Cancer Biochemical Recurrence Prediction using Machine Learning Analysis of Multiparametric Magnetic Resonance and Histopathology

Acronym: ProBCR

Portuguese Synopsis:
Neste estudo pretende-se utilizar métodos de aprendizagem automática para analisar imagens médicas de ressonância magnética multiparamétrica e de histopatologia digital, de pacientes com cancro da próstata tratados com prostatectomia radical. O objetivo é verificar se estes métodos são capazes de recolher informação das imagens que consiga prever a possibilidade de recorrência bioquímica de cancro da próstata. A capacidade de identificar indivíduos com essas características teria um impacto clínico importante.

English Synopsis:
The purpose of this study is to use machine learning methods to analyse medical images of multiparametric magnetic resonance and digital pathology, of prostate cancer patients treated with radical prostatectomy. It is intended to verify if these methods can extract relevant image information to predict the occurrence of biochemical recurrence of prostate cancer after surgical treatment. The ability to identify individuals with such characteristics would have an important clinical impact.

Type pf study:
Pilot retrospective study for patients already treated (with request for Informed Consent waiver).

Involved Institutes/Collaborations:
No other institutes will be involved or collaborations established for this study.

Funding:
No funding required for the execution of this project.

CR/CCC participation units/groups:
CR - Computational Clinical Imaging Group
CCC – Urology Unit
CCC - Radiology Department
CCC – Pathology Unit

Number of patients:
120-200

Study duration:
1 year

Research Protocol

Project Title

Prostate Cancer Biochemical Recurrence Prediction using Machine Learning Analysis of Multiparametric Magnetic Resonance and Histopathology

Investigators and Affiliation

Principal Investigator: Nickolas Papanikolaou, PhD[a]

Co-Investigators: Ana Gaivão, MD[b], António Beltran, MD, PhD[b]; Jorge Fonseca, MD[b]

Project Executors: Carolina Seabra, MSc Student[a]; Mónica Silva, MSc Student[a]

[a] Champalimaud Research, Lisbon, Portugal
[b] Champalimaud Foundation, Lisbon, Portugal

1 Background

1.1 Prostate Cancer and Biochemical Recurrence

Malignant neoplasms of the prostate, hereafter referred to as prostate cancer (PCa), usually originate in the glandular tissue. While these cancers are often indolent, there is a subset of men who are diagnosed with highly malignant prostate cancers associated with poor prognosis[1].

The disease poses a substantial public health burden worldwide: PCa is the most frequently diagnosed cancer among men in over-half of the countries in the world, and it is the leading cause of oncological death among men in 46 countries, with nearly 1.3 million new cases and 359000 associated deaths estimated for this year of 2018[2]. Selecting the optimal treatment for each patient is an important aspect in order to improve PCa clinical management.

Currently, PCa early clinical detection depends on the prostate specific antigen (PSA) serum level, as well as on the examination of multiparametric magnetic resonance imaging (mpMRI) that provides information on both morphological and physiological properties of tissues[3]. Nevertheless, its definitive diagnosis is based on histopathologic biopsy verification[4]. The pathologist attributes a prognostic predictor, the Gleason Score. This scoring system, developed in the 1960s, consists of two sub-grades: primary and secondary grades. The former is assigned to the dominant pattern of the tumour or the most common cell morphology, while the latter is assigned to the subordinate pattern. Each of the grades is defined on the scale from 1 to 5, according to the cellular and architectural appearance of recognisable glands, with lower grades corresponding to more normal prostate tissue[5].

In clinical practice, different treatment options are currently available for PCa patients, including active surveillance, adjuvant therapy, radiation therapy and radical prostatectomy (RP). The latter is, yet, the standard first-line curative procedure for the management of localised PCa[4,6–9], given the effectiveness of this therapeutic option for such patients[7,10,11]. It consists of surgically removing the prostate gland, the seminal vesicles, and surrounding tissue sufficient to ensure a negative surgical margins[12].

Besides being measured at the diagnostic stage, PSA serum level, an organ- but not cancer-specific biomarker, is also indispensably assessed in follow-up after a curative treatment, such as RP[13,14]. Its previous elevated value is expected to reach undetectable levels within 4 weeks after RP[15].

However, despite technical improvements in the surgical procedures for PCa treatments, there is a significant risk of cancer recurrence after therapy[7]. Of all patients undergoing RP, between 25% and 35%[7,16–20] present a rising detectable serum PSA level greater than 0,2 ng/ml[21] - a state known as biochemical recurrence (BCR).

Approximately two-thirds of BCRs occur within the first 2 years of surgery[22] and earlier BCR may be associated with increased risk of prostate cancer-specific mortality[23]. BCR is widely used as an end point to assess RP

efficacy, and it most likely represents the first sign of progression after surgery [24]. In some cases, its occurrence precedes cancer recurrence, representing, simultaneously, a marker of metastatic progression and PCa-specific mortality [6].

Thus, for men with PCa and candidates to receive curative treatment, such as RP, the risk of development of cancer recurrence after treatment is a main concern. Hence, an early BCR identification and treatment is of paramount importance to improve long-time survival.

1.2 Prediction of Biochemical Recurrence

One of the most acute needs in PCa management, nowadays, is higher precision in prediction of clinical outcomes for more effective decision-making [25].

Nomograms and probability graphs for BCR prediction following surgery have been constructed, such as the D'Amico et al. risk stratification scheme [26], the Stephenson et al. nomogram [27], and the Cancer of the Prostate Risk Assessment (CAPRA) score [28]. All of these models rely on commonly available clinical and histopathological variables, such as preoperative PSA, clinical stage, and biopsy Gleason score [6,29]. Although these nomograms have been internationally validated, only a few of them have predicted the probability of BCR with more than 70% accuracy [30–32].

Recent BCR prediction tools are incorporating mpMRI-derived variables to improve the outcome of the prediction models. This imaging modality has been investigated as promising way not only for detecting and staging the cancer, but also for risk stratification, since it is considered the most sensitive and most specific imaging technique to detect and characterise the clinical aggressiveness of the PCa tumours [7], which is related with BCR occurrence. On the other hand, including histopathologic digital images from the prostate resection might improve the accuracy of the tools which rely on biopsy information, since biopsy is not representative of the whole prostate lesion [33,34].

Given the current used methods, alternative approaches such as logistic regression models, support vector machines, classification and regression trees analysis, and artificial neural networks, are believed to further improve sensitivity, specificity and accuracy of the referred tools [35]. The combination of such computational approaches with data harvested from different imaging modalities might enhance BCR prediction for prostate cancer.

2 Project Aim

This study aims to develop classification models that can predict prostate cancer BCR.

One of the approaches will use imaging features that can be harvested from preoperative mpMRI data, while the other will be based on histopathologic prostate resection digital images. Both approaches will use preoperative clinical variables acquired in the current diagnostic protocol, and implement machine learning methods to predict BCR occurrence in patients who underwent radical prostatectomy for prostate cancer treatment.

For the former, the main objective would be to provide an initial evaluation, before a curative attempt, to patients diagnosed with PCa. Meaning, at the diagnostic time-point, and without the need of performing a biopsy, indirectly evaluate the biological aggressiveness of the patient's tumour - in the sense of relapsing or not after a RP surgery. The latter aims to take advantage of a novel way of analysing pathologic information, which is digital pathology. It presents several functionalities, such as the extraction of quantitative data, resulting from direct measurements like dimensions, number of pixels, or intensities, which cannot be accomplished with optical microscopy.

We hypothesise that there are different mpMR and histopathology digital image characteristics capable of differentiating between BCR-free and BCR patients, after undergoing radical prostatectomy.

3 Materials and Methods

3.1 Study Type

This pilot study will analyse retrospective data: preoperative clinical parameters, standard-of-care mpMR images routinely acquired at the diagnostic phase of prostate cancer management - without any changes to the clinical and imaging protocols -, and digital pathology images acquired from radical resection of the patients with PCa - without any changes to the clinical procedure.

3.2 Study Population

Regarding the histopathological imaging based approach, all patients who underwent radical prostatectomy at Champalimaud Clinical Center between 2016 and 2018 (200 patients) will be considered to be included in this study. As for the mpMR imaging based approach, all patients examined with preoperative prostate mpMRI and treated with radical prostatectomy at Champalimaud Clinical Centre, between 2016 and 2018 (approximately 120-150 patients) will be considered in the inclusion of this study. The imaging protocol used in these patients was the oncological standard of care in the management of prostate cancer.

Patients who received pre- or postoperative hormone or radiation therapy will be excluded from this study.

3.3 Patient Data Collection and Protection

The collection of the retrospective data will be performed by institutional radiologists, urologists and pathologists, reviewing, retrospectively, the clinical information of patients, and selecting the patients to be included on the analysis.

Regarding the mpMRI approach, the researchers and collaborators responsible for the project execution will be responsible for the retrieval of the patients' radiographic images from the institution Picture Archiving and Communication System (PACS) and de- identification of images and databases containing patient identifiable information.

For the histopathologic approach, the researcher responsible for the project execution will be responsible for the retrieval and scanning of the histopathologic digital images and de-identification of images and databases containing patient identifiable information.

For both procedures, a password protected encrypted file with the de-identification correspondence keys will be saved in a password protected workstation located within a room accessible only with the door key.

3.4 Data Processing, Analysis and Model Development

The collected data will be de-identified to ensure patient anonymity.

Imaging preprocessing methods will be explored for the enhancement of the image data. Machine learning algorithms will be applied to the set of data for the processing phase, since it is a powerful technique for recognising patterns on medical images[36]. This will allow the description of visible and invisible textures to the human eye in a numerical and objective way - known as radiomics features[37] -, followed by the model development to classify PCa patients based on their BCR occurrence. Lastly, model evaluation will be based on the outcome prediction performance.

4 Request for waiver of Informed Consent

The development of models to predict BCR will require the analysis of prostatectomy specimens digital images, mpMR images, as well as preoperative clinical information. The results of this study will be stronger by working with a larger number of data, and discarding the patients already imaged or followed in the past would delay the results. We will only use anonymised data from patients already treated, whose retrospective data analysis could benefit the treatment of future patients.

This pilot study might make possible the prediction of BCR, and this could improve the patients' treatment and/or prognosis control.

References

[1] F. Bray and L. A. Kiemeney, "Epidemiology of Prostate Cancer in Europe: Patterns, Trends and Determinants," in *Management of Prostate Cancer*, pp. 1–27, Cham: Springer International Publishing, 2017.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, sep 2018.

[3] N. B. Delongchamps, M. Rouanne, T. Flam, F. Beuvon, M. Liberatore, M. Zerbib, and F. Cornud, "Multiparametric magnetic resonance imaging for the detection and localization of prostate cancer: combination of T2-weighted, dynamic contrast-enhanced and diffusion-weighted imaging," *BJU International*, vol. 107, pp. 1411–1418, may 2011.

[4] N. Mottet, J. Bellmunt, M. Bolla, E. Briers, M. G. Cumberbatch, M. De Santis, N. Fossati, T. Gross, A. M. Henry, S. Joniau, T. B. Lam, M. D. Mason, V. B. Matveev, P. C. Moldovan, R. C. van den Bergh, T. Van den Broeck, H. G. van der Poel, T. H. van der Kwast, O. Rouvière, I. G. Schoots, T. Wiegel, and P. Cornford, "EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent," *European Urology*, vol. 71, pp. 618–629, apr 2017.

[5] J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, P. A. Humphrey, and Grading Committee, "The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma," *The American Journal of Surgical Pathology*, vol. 40, p. 1, oct 2015.

[6] G. Lughezzani, A. Briganti, P. I. Karakiewicz, M. W. Kattan, F. Montorsi, S. F. Shariat, and A. J. Vickers, "Predictive and Prognostic Models in Radical Prostatectomy Candidates: A Critical Analysis of the Literature," *European Urology*, vol. 58, pp. 687–700, nov 2010.

[7] N. Mottet, J. Bellmunt, E. Briers, M. Bolla, P. Cornford, M. De Santi, A. M. Henry, S. Joniau, T. B. Lam, V. B. Matveev, H. G. van der Poel, T. H. van der Kwast, O. Rouvière, T. Wiegel, T. van den Bergh, Roderick C.N. van den Broeck, and P. C. Moldovan, "EAU-ESTRO-SIOG Guidelines on Prostate Cancer," European Urology, 2017.

[8] H. Lepor, "Selecting Candidates for Radical Prostatectomy.," *Reviews in Urology*, vol. 2, no. 3, pp. 182–9, 2000.

[9] R. Tourinho-Barbosa, V. Srougi, I. Nunes-Silva, M. Baghdadi, G. Rembeyo, S. S. Eiffel, E. Barret, F. Rozet, M. Galiano, X. Cathelineau, and R. Sanchez-Salas, "Biochemical recurrence after radical prostatectomy: what does it mean?," *International Brazilian Journal of Urology*, vol. 44, no. 1, pp. 14–21, 2018.

[10] S. E. Eggener, P. T. Scardino, P. C. Walsh, M. Han, A. W. Partin, B. J. Trock, Z. Feng, D. P. Wood, J. A. Eastham, O. Yossepowitch, D. M. Rabah, M. W. Kattan, C. Yu, E. A. Klein, and A. J. Stephenson, "Predicting 15-Year Prostate Cancer Specific Mortality After Radical Prostatectomy," *The Journal of Urology*, vol. 185, pp. 869–875, mar 2011.

[11] A. B. Porcaro, A. Tafuri, M. Sebben, P. Corsi, T. Pocessali, M. Pirozzi, N. Amigoni, R. Rizzetto, A. Mariotto, D. Inverardi, M. Brunelli, R. Iacovelli, M. Romano, S. Siracusano, and W. Artibani, "Positive Association between Preoperative Total Testosterone Levels and Risk of Positive Surgical Margins by Prostate Cancer: Results in 476 Consecutive Patients Treated Only by Radical Prostatectomy.," *Urologia Internationalis*, vol. 101, no. 1, pp. 38–46, 2018.

[12] M. Nguyen-Nielsen and M. Borre, "Diagnostic and Therapeutic Strategies for Prostate Cancer," *Seminars in Nuclear Medicine*, vol. 46, pp. 484–490, nov 2016.

[13] D. Tilki, S. I. Kim, B. Hu, M. A. Dall'Era, and C. P. Evans, "Ultrasensitive Prostate Specific Antigen and its Role after Radical Prostatectomy: A Systematic Review," *The Journal of Urology*, vol. 193, pp. 1525–1531, may 2015.

[14] M. Kuriyama, M. C. Wang, C.-I. Lee, L. D. Papsidero, C. S. Killian, H. Inaji, N. H. Slack, T. Nishiura, G. P. Murphy, and T. M. Chu, "Use of Human Prostate-specific Antigen in Monitoring Prostate Cancer," *Cancer Research*, vol. 41, pp. 3874–6, oct 1981.

[15] A. J. Stephenson, M. W. Kattan, J. A. Eastham, Z. A. Dotan, F. J. Bianco, H. Lilja, and P. T. Scardino, "Defining Biochemical Recurrence of Prostate Cancer After Radical Prostatectomy: A Proposal for a Standardized Definition," *Journal of Clinical Oncology*, vol. 24, pp. 3973–8, aug 2006.

[16] J. F. Ward, M. L. Blute, J. Slezak, E. J. Bergstralh, and H. Zincke, "The Long-Term Clinical Impact of Biochemical Recurrence of Prostate Cancer 5 or More Years After Radical Prostatectomy," *The Journal of Urology*, vol. 170, pp. 1872–1876, nov 2003.

[17] S. J. Freedland, E. B. Humphreys, L. A. Mangold, M. Eisenberger, F. J. Dorey, P. C. Walsh, and A. W. Partin, "Risk of Prostate Cancer–Specific Mortality Following Biochemical Recurrence After Radical Prostatectomy," *JAMA*, vol. 294, p. 433, jul 2005.

[18] G. W. Hull, F. Rabbani, F. Abbas, T. M. Wheeler, M. W. Kattan, and P. T. Scardino, "Cancer control with radical prostatectomy alone in 1,000 consecutive patients.," *The Journal of urology*, vol. 167, pp. 528–34, feb 2002.

[19] C. L. Amling, M. L. Blute, E. J. Bergstralh, T. M. Seay, J. Slezak, and H. Zincke, "Long-term hazard of progression after radical prostatectomy for clinically localized prostate cancer: continued risk of biochemical failure after 5 years.," *The Journal of urology*, vol. 164, pp. 101–5, jul 2000.

[20] C. R. Pound, A. W. Partin, M. A. Eisenberger, D. W. Chan, J. D. Pearson, and P. C. Walsh, "Natural History of Progression After PSA Elevation Following Radical Prostatectomy," *JAMA*, vol. 281, p. 1591, may 1999.

[21] P. Cornford, J. Bellmunt, M. Bolla, E. Briers, M. De Santis, T. Gross, A. M. Henry, S. Joniau, T. B. Lam, M. D. Mason, H. G. van der Poel, T. H. van der Kwast, O. Rouvière, T. Wiegel, and N. Mottet, "EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer," *European Urology*, vol. 71, pp. 630–642, apr 2017.

[22] J. Walz, F. K.-H. Chun, E. A. Klein, A. Reuther, F. Saad, M. Graefen, H. Huland, and P. I. Karakiewicz, "Nomogram Predicting the Probability of Early Recurrence After Radical Prostatectomy for Prostate Cancer," *The Journal of Urology*, vol. 181, pp. 601–608, feb 2009.

[23] S. J. Freedland, E. B. Humphreys, L. A. Mangold, M. Eisenberger, and A. W. Partin, "Time to Prostate Specific Antigen Recurrence After Radical Prostatectomy and Risk of Prostate Cancer Specific Mortality," *The Journal of Urology*, vol. 176, pp. 1404–1408, oct 2006.

[24] L. Budäus, J. Schiffmann, M. Graefen, H. Huland, P. Tennstedt, A. Siegmann, D. Böhmer, V. Budach, D. Bartkowiak, and T. Wiegel, "Defining biochemical recurrence after radical prostatectomy and timing of early salvage radiotherapy," *Strahlentherapie und Onkologie*, vol. 193, pp. 692–699, sep 2017.

[25] S. F. Shariat, M. W. Kattan, A. J. Vickers, P. I. Karakiewicz, and P. T. Scardino, "Critical review of prostate cancer predictive tools," *Future Oncology*, vol. 5, pp. 1555–1584, dec 2009.

[26] A. V. D'Amico, R. Whittington, S. B. Malkowicz, D. Schultz, K. Blank, G. A. Broderick, J. E. Tomaszewski, A. A. Renshaw, I. Kaplan, C. J. Beard, and A. Wein, "Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer," *JAMA*, vol. 280, p. 969, sep 1998.

[27] A. J. Stephenson, P. T. Scardino, J. A. Eastham, F. J. Bianco, Z. A. Dotan, P. A. Fearn, and M. W. Kattan, "Preoperative Nomogram Predicting the 10-Year Probability of Prostate Cancer Recurrence After Radical Prostatectomy," *JNCI: Journal of the National Cancer Institute*, vol. 98, pp. 715–717, may 2006.

[28] M. R. Cooperberg, D. J. Pasta, E. P. Elkin, M. S. Litwin, D. M. Latini, J. Du Chane, and P. R. Carroll, "The University of California, San Francisco Cancer of the Prostate Risk Assessment score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy.," *The Journal of Urology*, vol. 173, pp. 1938–42, jun 2005.

[29] K. Nishida, S. Yuen, K. Kamoi, K. Yamada, K. Akazawa, H. Ito, K. Okihara, A. Kawauchi, T. Miki, and T. Nishimura, "Incremental value of T2-weighted and diffusion-weighted MRI for prediction of biochemical recurrence after radical prostatectomy in clinically localized prostate cancer," *Acta Radiologica*, vol. 52, pp. 120–126, feb 2011.

[30] V. Poulakis, U. Witzsch, R. de Vries, V. Emmerlich, M. Meves, H.-M. Altmannsberger, and E. Becht, "Preoperative neural network using combined magnetic resonance imaging variables, prostate-specific antigen, and gleason score for predicting prostate cancer biochemical recurrence after radical prostatectomy," *Urology*, vol. 64, pp. 1165–1170, dec 2004.

[31] G. Lughezzani, L. Budäus, H. Isbarn, M. Sun, P. Perrotte, A. Haese, F. K. Chun, T. Schlomm, T. Steuber, H. Heinzer, H. Huland, F. Montorsi, M. Graefen, and P. I. Karakiewicz, "Head-to-Head Comparison of the Three Most Commonly Used Preoperative Models for Prediction of Biochemical Recurrence After Radical Prostatectomy," *European Urology*, vol. 57, pp. 562–568, apr 2010.

[32] J. Morote, J. del Amo, A. Borque, E. Ars, C. Hernández, F. Herranz, A. Arruza, R. Llarena, J. Planas, M. J. Viso, J. Palou, C. X. Raventós, D. Tejedor, M. Artieda, L. Simón, A. Martínez, and L. A. Rioja, "Improved Prediction of Biochemical Recurrence After Radical Prostatectomy by Genetic Polymorphisms," *The Journal of Urology*, vol. 184, pp. 506–511, aug 2010.

[33] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, and N. Rajpoot, "Histopathological Image Analysis: A Review NIH Public Access," *IEEE Rev Biomed Eng*, vol. 2, pp. 147–171, 2009.

[34] J. L. Fine, D. M. Grzybicki, R. Silowash Bs, J. Ho, J. R. Gilbertson, L. A. Ma, R. W. Ma, A. V. Parwani, S. I. Bastacky, J. I. Epstein, and D. M. Jukic, "Evaluation of whole slide image immunohistochemistry interpretation in challenging prostate needle biopsies B," *Human Pathology*, vol. 39, pp. 564–572, 2008.

[35] M. W. Kattan, A. M. Stapleton, T. M. Wheeler, and P. T. Scardino, "Evaluation of a nomogram used to predict the pathologic stage of clinically localized prostate carcinoma.," *Cancer*, vol. 79, pp. 528–37, feb 1997.

[36] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine Learning for Medical Imaging," *Radiographics: a review publication of the Radiological Society of North America, Inc*, vol. 37, no. 2, pp. 505–515, 2017.

[37] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, pp. e104–e107, nov 2017.

Informed Consent Form

This informed consent form is directed to patients who underwent radical prostatectomy.

The title of the research project is *"Prostate Cancer Biochemical Recurrence Prediction using Machine Learning Analysis of Multiparametric Magnetic Resonance and Histopathology"*.

This is a national project developed by Champalimaud Foundation.

Principal Investigator

Nickolas Papanikolaou

Introduction and Purpose

The Urology Unit and Computational Clinical Imaging group together with the Pathology and Radiology Clinical Service from Champalimaud Foundation, have a research project on biochemical recurrence following radical prostatectomy. The main purpose is to use multiparametric magnetic ressonace (mpMR) imaging and histopathology to predict prostate cancer biochemical recurrence after radical prostatectomy surgical treatment.

After hearing the procedure information, we present a brief description about the study. Please read carefully the following information. After reading, if you have any doubts, do not hesitate to ask them to your physician so they can be clarified.

Type of intervention
This study did not change the clinical and therapeutic procedure previously defined. We solely ask you the permission for the researcher to use your mp-MR and histopathologic images to apply new analysis methods.

Patient selection
All patients with prostate cancer who were examined through mp-MR imaging and underwent radical prostatectomy, at the Champalimaud Foundation may be included in the study. In case you received any type of pre- or postoperative treatment, your enrolment in this study should be disregarded.

Voluntary participation
The participation in this study is voluntary. The patient is free to choose whether or not to participate. The patient's choice did not change the quality of the treatment he received. In case you want to participate, you are free to dismiss at any moment, needing only to inform your urologist.

Procedure
The participation in this study did not involve additional procedures. The follow-up of the patient was done according with the institution's standard of care of oncological patients.

Risks
No risks were defined for the participation on this study.

Benefits
The results of this study can potentially improve the treatment of prostate cancer patients who underwent radical prostatectomy.

Compensations
Your participation in this study will not have any additional costs for you. No compensation, either monetary or other, will be given to you for your participation in this study.

Confidentiality
All the data collected during this study will be confidential and protected. In case you participate, a unique identifier will be assigned to you and only your physician will know that this unique identifier corresponds to you.

Sharing of the results
The results of this study, ensuring the correct de-identification, can be shared on scientific meetings and published in scientific journals.

Right to refuse or dismiss
Your participation in this study is your choice. Whether you chose or not to participate in this study the quality and the treatments that you receive will not change. In case you decide to dismiss, you can do it at any moment by contacting your urologist.

Contacts
In case you have any questions in the future, either during or after this study, please contact the urologist in charge:

Dr.	**Contact**
_____	
Name of the patient	
_____	_____
Signature of the patient	Date (dd/mm/yyyy)
_____	
Name of the urologist	
_____	_____
Signature of the urologist	Date (dd/mm/yyyy)

Formulário de Consentimento Informado

Este formulário de consentimento informado é dirigido a pacientes com cancro da próstata tratados através de prostatectomia radical.

O título do projeto de investigação é *"Prostate Cancer Biochemical Recurrence Prediction using Machine Learning Analysis of Multiparametric Resonance Imaging and Histopathology"*.

Este projeto é nacional e foi desenvolvido pela Fundação Champalimaud.

Investigator Principal

Nickolas Papanikolaou

Introdução e Propósito de Estudo

A Unidade de Urologia e o grupo *Computational Clinical Imaging*, em conjunto com os Serviços Clínicos de Patologia e Radiologia da Fundação Champalimaud, estão a realizar um estudo em pacientes com neeoplasia maligna da próstata. O objetivo principal é desenvolver um modelo, baseado em imagens de ressonância magnética multiparamétrica (RM-mp) e histopatologia da próstata, que permita prever recorrência bioquímica, após tratamento com prostatectomia radical.

Após a informação sobre o processo lhe ter sido prestada oralmente, apresenta-se em seguida um resumo escrito das características deste estudo. Por favor leia com atenção esta informação. Se, após ler esta informação, tiver quaisquer dúvidas, por favor coloque-as ao seu médico para que possam ser esclarecidas.

Tipo de intervenção

O estudo não alterou os planos clínico e terapêutico, previamente estabelecidos. Pedimos-lhe, somente, permissão para que os investigadores possam avaliar novos métodos de análise de imagem referentes às imagens de RM-mp e de histopatologia.

Seleção de participantes

Todos os doentes com cancro da próstata que foram examinados através de RM-mp e tratados com prostatectomia radical na Fundação Champalimaud, poderão ser incluídos no estudo. Caso tenha realizado outro tipo de tratamento pré- ou pós-operatório, não deverá ser incluído neste estudo.

Participação voluntária

A participação neste estudo é totalmente voluntária. A escolha de participar ou não neste estudo cabe unicamente ao doente. O facto de escolher ou não participar neste estudo não vai alterar a qualidade dos serviços e tratamentos que recebe. Caso escolha participar, o doente poderá desistir a qualquer momento, bastando para isso informar o seu médico urologista.

Procedimentos

Não foram realizados procedimentos adicionais como resultado da participação no estudo. O seguimento do doente enquadrou-se no seguimento normal de doentes oncológicos realizado pela Instituição de origem.

Riscos

Não foram definidos riscos de participação neste estudo.

Benefícios
Os resultados do estudo têm o potencial de melhorar o tratamento de doentes com cancro da próstata tratados através de cirurgia radical da próstata.

Compensações
A sua participação neste estudo não terá quaisquer custos adicionais para si. Não lhe será dada nenhuma compensação, monetária ou outra, pela sua participação neste estudo.

Confidencialidade
Todos os dados recolhidos durante este estudo são mantidos estritamente confidenciais, em suporte informático protegido. Se decidir participar neste estudo, ser-lhe-á atribuído um número de identificação único. Apenas o seu médico saberá que esse número único corresponde a si.

Partilha de resultados
Os resultados do estudo, devidamente anonimizados, serão partilhados em reuniões científicas e publicados em revistas científicas.

Direito a recusar ou desistir
A sua participação neste estudo é uma escolha exclusivamente sua. O facto de escolher ou não participar neste estudo não vai alterar em nada a qualidade dos serviços e tratamentos que recebe. Caso escolha participar poderá desistir a qualquer momento, bastando para isso informar o seu médico.

Contatos
Caso lhe surjam quaisquer questões futuras, seja durante a sua participação neste estudo ou depois de este ter terminado, por favor contacte o médico urologista responsável:

Dr. **Contacto**

_____
Nome do paciente

_____ _____
Assinatura do paciente Data (dd/mm/yyyy)

_____
Nome do médico

_____ _____
Assinatura do médico Data (dd/mm/yyyy)